Programa de Doctorat en ciències

Escola de Doctorat de la Universitat Jaume I

DESARROLLO DE METODOLOGÍAS METABOLÓMICAS NO DIRIGIDAS BASADAS EN CROMATOGRAFIA DE LÍQUIDOS DE ULTRA ALTO RENDIMIENTO ACOPLADO A ESPECTROMETRÍA DE MASAS DE ALTA RESOLUCIÓN EN EL CAMPO DE SEGURIDAD ALIMENTARIA, SANIDAD Y NUTRICION

Memòria presentada per Rubén Gil Solsona per a optar al grau de doctor/a per la Universitat Jaume I

Fdo. Rubén Gil Solsona

Fdo. Dr. Juan Vicente Sancho Llopis

Fdo. Dr. Jaume Pérez Sánchez

Castelló de la Plana, Octubre 2018

Finançament Rebut / Financiamiento Recibido

Agències finançadores del doctorand/a

· PROMETEO II/2014/023

· Research Unit of Marine Ecotoxicology (IATS-CSIC; IUPA-UJI).

· SCORE-COST action ES1307. Short Term Scientific Mission (STSM) grant (COST-STSM-ES1307-150916-080342

Agències finançadores del projecte de recerca o dels recursos materials específics del grup de recerca. /

· NPS-Euronet

· Ministerio Español de Economia y Competititvidad

· Generalitat Valenciana

"Reports that say that something hasn't happened are always interesting to me, because as we know, there are known knowns; there are things we know we know. We also know there are known unknowns; that is to say we know there are some things we do not know. But there are also unknown unknowns – the ones we don't know we don't know. And if one looks throughout the history of our country and other free countries, it is the latter category that tend to be the difficult ones"

Donald Rumsfeld, United States Secretary, 12th February 2002

Interview about the lack of evidences linking the government of Iraq with the supply of weapons of mass destruction.

"~~Los informes que dicen que algo que no ha pasado resulta siempre interesante, porque como ya sabemos, hay conocidos que conocemos; hay cosas que sabemos que sabemos~~. Nosotros también sabemos que hay cosas conocidas que desconocemos; o lo que es lo mismo, sabemos que hay algunas cosas que no conocemos. Sin embargo hay también desconocidos desconocidos – que son los que no sabemos que no sabemos. Y si uno mira a través de la ~~historia de nuestro país y de otros países libres~~ **metabolómica**, esta es la última categoría y la que suele ser más dificultosa"

Estudiante anónimo de doctorado

Conversación sobre cómo definiría la metabolómica no dirigida

AGRADECIMIENTOS

Como todo en esta vida, el periodo de doctorado también tiene un final. Y como todo en esta vida, los finales siempre resultan ser la parte más dura. Sin embargo, no es siempre llegar a la meta la parte más importante, y por ello hay que acordarse siempre del camino que se ha recorrido para alcanzarlo. Por este motivo, ha llegado el momento de dar las gracias.

Por comenzar por el principio, muchas gracias a mis directores de tesis. A Juanvi, por aceptarme cuando era un alumno de Análisis de Alimentos que quería probar el mundo de la investigación. Las reuniones y los viajes contigo siempre aportaban cosas nuevas, aunque la *Ley de Murphy* dijera lo contrario... Gracias también a Jaume, con quien he aprendido que cualquier trabajo funciona mejor si se ve desde diferentes puntos de vista y quien me ha enseñado la mayoría de mis escasos conocimientos en biología.

Gracias a mis compañeros del IUPA por acompañarme durante todo el trayecto, no solo con su conocimientos sino con su compañía y apoyo.

Quiero agradecer en especial a Montse, por todo el apoyo que me ha brindado y de quien he aprendido muchísimo más de lo que yo he aportado. Por eso un gracias se queda mucho más que corto.

También quiero agradecer a Carlos todo el tiempo que hemos pasado juntos, compartiendo muchísimo (incluso habitaciones de hotel...). Todo el tiempo que hemos pasado juntos me ha ayudado más de lo que puede pensar.

A Edu y Clara, con quienes he compartido mucho más que ratos de risas y cervezas... Sus siempre valiosos consejos han estado disponibles cuando los he necesitado.

Señor Duván! Gracias a tu visita aprendí mucha más metabolómica de la que te enseñé. Siempre quedará Magdalena para recordar tu paso por Castellón.

A María, por todo el apoyo que recibí por su parte cuando comenzó mi aventura con la alta resolución, y que todavía continúa a día de hoy, por lo que estaré agradecido por siempre. A Tania, con quien aprendí muuucha informática, y quien ha estado ahí cuando la he necesitado.

A Leticia, por todas las horas que ha dedicado a llevar adelante con ganas y entusiasmo los trabajos que tenemos en común.

Gracias a Robert y Mercedes, quienes a pesar de habitar un despacho alejado del mundo siempre han estado ahí en los momentos oportunos.

Thanks to Olivier, Pierre and Céline for allowing me to stay in their office and their university. I want to thanks Olivier the moments of laughs that everyone needs in

these moments far away from home. Agradecer también a Juan por el apoyo que me ha dado en mi estancia y por hacerme abrir los ojos en los momentos necesarios.

Quiero además agradecer a mis amigos por todo el apoyo que he recibido durante mi vida, ya que han estado ahí antes y lo estarán después de la tesis. Necesito hacer mención especial a mi *hermano* Zarach, con quien la vida tiene otro sentido y con quien puedes contar en los buenos y malos momentos, siempre con un sabio consejo de metalero. También a Antonio, David, Victor, Alba, Cristina... con quienes siempre puedes contar, a pesar de lo descuidados que los tengo...

Quiero agradecer a toda mi familia el apoyo, aunque quiero hacer mención especial a mis padres por todo el tiempo que han dedicado a permitirme llevar a cabo mis sueños. Padres no hay más que dos, pero tengo suerte de tener los que tengo. Tampoco querría olvidar a mi hermano, de quien podría decir mil cosas buenas y muy pocas malas. Siempre ha estado ahí en el momento justo, lo cual resulta difícil de tener hoy en día.

Finalmente, como buen comilón que soy, guardo lo mejor para el final. Muchas gracias a Lorena, quien me ha acompañado durante este largo y sinuoso trayecto. Su apoyo incondicional y su compañía me han hecho levantarme en ocasiones donde no hubiera resultado fácil hacerlo sin ayuda.

Ella sabe lo mucho que la quiero, aunque siempre viene bien dejarlo plasmado en este momento tan especial de mi vida, para que, al igual que el resto, puedan echar la

vista atrás, leer estas líneas y recordar que, a pesar de que la meta se vislumbra cada vez más cercana, hemos pasado por otros periodos que nos han convertido en lo que somos, y que es en realidad nuestro objetivo.

Resumen

Las nuevas oportunidades que proporciona el avance de la ciencia, actualizando las metodologías analíticas, han propiciado la aparición de instrumentos de espectrometría de masas de mayor resolución, acoplamientos de cromatografía más robustos y fiables e incluso aportes como la bioinformática; o la combinación de todas ellas con la aparición de la metabolómica. Esta aproximación o flujo de trabajo, que inicialmente apareció como un complemento de otras tecnologías ómicas (genómica, transcriptómica, proteómica), proporciona una manera de trabajar diametralmente opuesta a las aproximaciones convencionales. Inicialmente la ciencia proponía hipótesis que debían ser contrastadas para aceptarlas o no, la metabolómica no dirigida permite establecer hipótesis "*conociendo la respuesta*". Esto es así debido a que la hipótesis se plantea tras obtener una visión global del problema a abordar y de las diferencias en los compuestos (posibles soluciones), por lo que generalmente la hipótesis coincide con la respuesta.

Esta aproximación ha sido aplicada desde su aparición en los años 2000 en muy diferentes campos, como diagnóstico de enfermedades, autentificación de alimentos o estudio de metabolismo de drogas, tóxicos, etc. Desde entonces se han utilizado diferentes plataformas analíticas, entre las que se encuentran la resonancia magnética nuclear (RMN), la espectrometría de masas (MS) o diferentes espectroscopias. Cada una de ellas presenta ventajas e inconvenientes, aunque se podría destacar el uso de RMN y MS por encima del resto, siendo el punto fuerte de RMN su poder de elucidación y universalidad, mientras que MS tiene muchísima mejor sensibilidad. Además el acoplamiento de esta última a separaciones cromatográficas, tanto de líquidos como de gases, y la aparición de librerías y bases de datos online ha puesto el uso de MS por delante de RMN en este campo.

En la presente Tesis Doctoral se ha evaluado, desde la plataforma metabolómica basada en instrumentos de cromatografía de líquidos de ultra alta resolución (UHPLC) acoplada a espectrometría de masas de alta resolución (HRMS), diferentes aplicaciones que permitan obtener respuestas analíticas de un modo fiable para diferentes problemas que han surgido en la sociedad. A pesar de que todo el trabajo desarrollado en la presente tesis se basa en la multiplataforma UHPLC- HRMS – metabolómica, se observan tres capítulos diferenciados dentro de ella:

Resumen

En la primera parte (Capítulo II), se muestra la aplicación de la metabolómica en estudios de nutrición. Los experimentos llevados a cabo con doradas (*Sparus Aurata*), tenían por objetivo observar los compuestos que se veían alterados en mayor medida en sangre derivados de los desafíos nutricionales a las que eran sometidas. En un primer artículo, a modo de prueba de concepto y potencial de la espectrometría de masas en este campo para obtener marcadores muy alterados, las doradas se sometieron a un prolongado periodo de ayuno. En el experimento se observó cómo sus parámetros biométricos se vieron realmente alterados, y la espectrometría de masas proporcionó un número extremadamente elevado de compuestos perturbados. De entre ellos, los que habían cambiado de una manera más abrumadora fueron seleccionados para su elucidación y, tras esta, su funcionalidad biológica fue establecida. Con el objetivo de esclarecer y asegurar la vía metabólica alterada, compuestos relacionados se buscaron en los datos ya adquiridos, gracias a la funcionalidad del QTof trabajando en modo MS^E, donde la información obtenida es de espectro completo con y sin fragmentación. En un segundo experimento, se modificó la composición de las dietas en los peces (de una cantidad alta de lípidos y proteínas animales a otra con mayor contenido vegetal). El objetivo era comprobar que las diferencias observadas entre los grupos se debían únicamente a una modificación de los nutrientes y no a una malnutrición de los animales. Se comprobó también que los perfiles de micronutrientes se mantenían en un nivel correcto, además de elucidar una gran cantidad de compuestos alterados por medio de la utilización de cromatografía de líquidos de ultra alta resolución (UHPLC) acoplada a espectrometría de masas de alta resolución (HRMS) apoyado en librerías de espectros online.

En la segunda parte (Capítulo III) se trabajó en la aplicación de la metabolómica con el objetivo de crear modelos estadísticos de autentificación de alimentos en base a la información obtenida por los instrumentos de UHPLC-HRMS por cuadrupolo tiempo de vuelo (QTof). Se llevó a cabo un estudio con diferentes tipos de cromatografía para obtener estos marcadores discriminantes de aceites de oliva españoles por zona geográfica, de almendras en país de origen y variedad de las españolas y de café colombiano por zona geográfica. Estos compuestos fueron aislados por la estadística multivariante dirigida (PLS-DA (filtrado por VIP), OPLS-DA (filtrado por

P[corr]) y posteriormente se elucidaron algunos de ellos, gracias a la información proporcionada por los experimentos de espectrometría de masas en tándem (MS/MS). Estos marcadores se utilizaron posteriormente para crear modelos de discriminación, que fueron además validados con muestras de siguientes temporadas, añadiendo robustez a su uso con este objetivo.

En la tercera parte (capítulo IV) se trabajó en el desarrollo de un nuevo modelo para afrontar el problema que ha aparecido con las nuevas drogas psicoactivas. Los *cannabinoides* sintéticos (SCs), los cuales se modifican constantemente evadiendo problemas con la legalidad, se rocían sobre bases compuestas por hierbas con teóricas propiedades psicotrópicas, en mayor o menor medida, para ser vendidas así como sustitutos de la marihuana. Para combatir esta constante actualización de los SCs se ha propuesto un método para obtener biomarcadores de estas mezclas de hierbas para controlar así el uso indiscriminado de las nuevas sustancias desconocidas. En una primera aproximación (artículo científico VI) se probó como la estrategia realmente funciona y puede ser útil obteniendo dos biomarcadores de un grupo tan heterogéneo de hierbas en la saliva de voluntarios que las fumaron mezcladas con tabaco. Estos marcadores, a su vez, aparecen con una relación diferente en cigarros compuestos únicamente por tabaco y en los mezclados con las hierbas. Tras esta constatación, se establecieron cuatro posibles muestras no invasivas a controlar, el humo de los cigarros, el exhalado pulmonar contaminado por este humo, la saliva también contaminada por éste y finalmente la orina. Incialmente, se seleccionó el humo producido en la combustión como primera matriz a investigar. En este artículo, adicionalmente, se evalúa como los nuevos instrumentos que han aparecido combinado las celdas de movilidad iónica con UHPLC-HRMS, proporcionan un poder de confirmación superior a los tradicionales, además de mayor confianza en la elucidación con el uso conjunto de herramientas de predicción de tiempo de retención y sección de colisión transversal. De este modo se obtuvieron dos marcadores que permiten diferenciar el humo de tabaco del humo de estas hierbas, alcanzándose una elucidación bastante fiable de los mismos.

Summary

New opportunities provided by science progress, which updates analytical methodologies, have promoted mass spectrometer instruments with higher resolution, more robust and reliable couplings between them and chromatographic techniques and even other new resources like bioinformatics or, boosted by the joining of all of them, metabolomics. This approach or workflow, which appeared as other omic techniques complement (genomics, transcriptomics, proteomics), provides an orthogonal way to face scientific challenges. Until these techniques appeared, science hypothesize about feasible solutions, which should be contrasted to be accepted or discarded. However, untargeted metabolomics allows scientists to hypothesize "*knowing the answer*". It can be done as metabolomics provides a general overview about the problem, providing differences between compounds (solutions) before hypothesize, so generally this hypothesis match with the solution.

This approach have been applied, since it appeared in the 2000's, in many different fields, as disease research, food authenticity or *in-vivo* metabolism for drugs, toxicants,…. Since then, many different analytical platforms have been employed, such as nuclear magnetic resonance (NMR), mass spectrometry (MS) or different spectroscopic techniques among others. Each platform provides some benefits and drawbacks to the approach, but the use of NMR and MS have outweighed the rest. The main advantage of NMR is its universality and elucidation power, being extremely less sensitive compared to MS. Furthermore, the easy coupling of MS platforms to liquid and gas chromatographic techniques together with the appearance of libraries and online databases have boosted the use of MS-based metabolomics, overtaking NMR-based metabolomics.

In the present Doctoral Thesis, based on ultra-high performance liquid chromatography (UHPLC) coupled to high resolution mass spectrometry (HRMS) instruments, different applications have been evaluated. The main objective was to obtain analytical responses in a reliable way to solve different social problems. Despite the experiments have been performed with this multiplatform, three different chapters can be observed in this work:

In the first part (Chapter II), metabolomics applications in nutrition studies are shown. These experiments, which have been carried out with Gilthead Sea Bream (*Sparus Aurata*), had the objective to measure altered metabolites in fish blood derived from nutritional challenges. In the first scientific article, as a proof-of-concept for showing the potential of mass spectrometry in this field to obtain highly altered biomarkers, fish were fasted during a long period of time. In the experiment biometric data was altered pointing out the magnitude of the change. MS provided a large amount of altered metabolites. Among them, the most highly altered metabolites were selected for elucidation. Elucidated compounds were then linked with a metabolic pathway. After that, related compounds were searched in order to confirm the altered pathway. This was possible due to the MS^E acquisition mode of available QTof analyzer, which collect full scan information at different collision energies, where HRMS spectra, with and without fragmentation, is obtained. In a second experiment, fish feed composition was modified from a high amount of fish protein and fish lipids to a higher vegetal composition. The objective was to prove that the observed differences between groups were mainly related to nutrient composition of these diets, discarding fish malnutrition. Micronutrient profile was also measured in order to ensure that no difference was achieved between fish fed with these diets. Additionally, a high amount of altered compounds were tentatively elucidated thanks to the UHPLC-HRMS platform, supported by online spectral databases and libraries.

In a second part (Chapter III), UHPLC-HRMS-based metabolomics was employed to perform authentication statistical models in food. Different chromatographic modes were employed to perform a wide-polarity coverage. The objective was to point out biomarkers which allows to distinguish between extra virgin olive oil (EVOO) from Spain, almonds from Spain and other countries, as well as different Spanish varieties and Colombian coffee cultured in different cultivars. These compounds were isolated and highlighted by means of multivariate statistical analysis (VIP filtering in PLS-DA or p[corr] filtering in OPLS-DA). Then, some of them were tentatively elucidated, with the information provided by tandem mass spectrometry experiments (MS/MS). These markers were employed to create discrimination models, which were validated with second season samples, ensuring robustness to the models, developed for using them as an authentication analytical tool.

Summary

In the third part (Chapter IV), metabolomics was applied to face out the new psychoactive substances (NPS) problem. Synthetic cannabinoids (SCs), which structure are being constantly modified to avoid legal problems, are sprayed over herbal blends which are supposed to be psychoactive, to be sold as marijuana substitutes. In order to cope with this constant update in SCs, a new methodology have been proposed for biomarkers discovering in these herbal base in order to control indirectly the indiscriminate use of these unknown substances. In a first approach (scientific article VI) the platform was tested to observe if these metabolites can be highlighted from a heterogeneous herbal group in saliva of volunteers smoking these herbs mixed with tobacco. Two biomarkers were highlighted and their ratio was employed to differentiate between them and, in turn, they had different ratio in tobacco samples and in herb-tobacco ones. After this evidence, four different non-invasive samples were thought to be employed for this purpose, smoke directly obtained from the cigarettes, exhaled breath contaminated by this smoke, saliva with this compounds and also urine. For this reason, smoke was firstly selected. In this article, additionally to the biomarkers discovery objective, the usefulness of acquiring ion mobility information was proved. These new instruments, combining ion mobility cells with UHPLC-HRMS instruments provides higher confirmation power, related to conventional instruments. They also provide extra confidence in elucidation purposes linking to both retention time and collision cross section prediction tools. In this way, two markers were highlighted to allow the differentiation between tobacco smoke and herb smoke, achieving reliable elucidation confidence.

INDICE GENERAL

Objetivos

Esta tesis doctoral ha sido desarrollada mediante la colaboración del grupo de investigación del Instituto Universitario de Plaguicidas y Agua con otros grupos de investigación nacionales y/o internaciones, que han resultado esenciales para el desarrollo de un trabajo tan completo y multidisciplinar. El capítulo II ha sido llevado a cabo gracias a la colaboración con el grupo de Nutrigenómica del Dr. Pérez Sánchez, del Instituto de Acuicultura de Torre la Sal (CSIC), codirector de esta tesis doctoral. El capítulo III ha sido desarrollado por el grupo de investigación español, con mención especial al grupo del Dr. Peñuela, con quien se ha llevado a cabo una estrecha colaboración en el estudio de asignación de denominación de origen de café de Colombia. Por último, el capítulo IV se ha desarrollado gracias a una colaboración internacional con el Dr. Olivier Delémont y su grupo de investigación en ciencias forenses de la Universidad de Lausana (Suiza), donde se desarrolló la estancia investigadora de tres meses del autor de la presente tesis doctoral. De todas y cada una de ellas surge el objetivo general que se expone:

- Evaluación de las técnicas analíticas basadas en el acoplamiento de la cromatografía de líquidos de ultra alto rendimiento (UHPLC) con la espectrometría de masas de alta resolución (HRMS) que permitan obtener información sobre la composición de diferentes tipos de matriz (alimentos, biofluidos,…) con el objetivo de extraer y elucidar los compuestos discriminantes más relevantes para cada experimento en concreto.

A raíz de éste, se pueden extraer diferentes objetivos específicos para cada uno de los trabajos desarrollados:

- Obtener una amplia visión de los compuestos presentes en los alimentos analizados en esta tesis (aceite de oliva, almendras y café verde) para seleccionar los que mejor separen los alimentos por su zona geográfica o variedad. A partir de esta, elucidar gracias a la potencia de la espectrometría de masas de alta resolución y los experimentos de masas en tándem, una lista de compuestos seleccionados que permitan crear un modelo estadístico de discriminación.

- Obtener un amplio conocimiento sobre los biomarcadores que pueden estar relacionados en diferentes pruebas nutricionales a las que se han sometido las doradas

(Sparus Aurata), bien sea a la falta de alimento o a la variación de la composición de las dietas.

- Evaluar las posibilidades y requerimientos que tiene el uso de una cuarta dimensión (movilidad iónica) en el tratamiento de datos y en el trabajo de elucidación de los compuestos químicos en comparación con un instrumento sin esta característica.

- Obtener biomarcadores indirectos de consumo de sustancias estupefacientes (nuevos cannabinoides sintéticos) por medio del estudio de las mezclas de hierbas utilizadas como base de estos productos en la saliva de fumadores de estas sustancias.

Plan de trabajo

Para alcanzar los objetivos indicados anteriormente, el plan de trabajo que se trazó para todos los experimentos fue el siguiente:

- Estudio de las condiciones de extracción para las diferentes matrices, con el objetivo de perder la mínima cantidad de información posible en este punto del flujo de trabajo pero también de obtener un amplio rango de compuestos.

- Inyección de las muestras de manera aleatoria para evitar así diferencias generadas por el orden de inyección (errores sistemáticos).

- Estudio y optimización de los parámetros de procesamiento de datos (*peak picking, retention time alignment, normalization*,..).

- Estudio de la varianza de los datos por medio de técnicas no dirigidas (Análisis de Componentes Principales, PCA) con el objetivo de controlar la variación de los sets de muestras por medio de un "patrón externo" o muestra QC, que se trata en todos los casos de una mezcla a partes iguales de todas las muestras del set, con el objetivo de tener una muestra que pueda ser representativa de todas a la vez que controle la deriva instrumental.

- Desarrollo de modelos estadísticos, bien univariantes (ANOVA) o multivariantes (PLS-DA, OPLS-DA, DD-SIMCA), que dependerá del objetivo de cada estudio en concreto.

- Elucidación del máximo número de compuestos posible en cada caso.
- Confirmación, en caso de ser posible, de las identificaciones tentativas mediante la comparación con patrones estándar obtenidos comercialmente.

Objectives

This PhD thesis have been developed under the collaboration of Research Institute for Pesticides and Water (IUPA) research group with other national and/or international research groups, which have been essential to perform a complete and interdisciplinary work. Chapter II have been carried out thanks to the collaboration with Dr. Pérez Sánchez Nutrigenomics research group, from Aquaculture Institute of Torre la Sal (CSIC), co-director of this PhD thesis. Chapter III have been developed by IUPA research group, with special mention to Dr. Peñuela research group, a close collaboration have been maintained in green Colombia coffe PDO assessment. Finally, chapter IV have been performed thanks to the international collaboration with Dr. Olivier Delémont and its forensic sciences research group of Lausanne University (Switzerland), where the PhD author stayed for three months during a Short Term Scientific Mission. In all these collaborations, the general objective of this PhD was as follows:

- Evaluation of analytical techniques based on the combination of ultra-high performance liquid chromatography (UHPLC) coupled to high resolution mass spectrometry (HRMS) which allows to obtain information about matrix compositions (food, biofluids,…) with the main goal to extract and elucidate the most relevant discriminant compounds in each experiment.

Consequently to this objective, it is possible to extract other specific objectives in each developed work:

- To obtain a wide coverage of compounds present in analyzed food in this PhD work (olive oil, almonds and green coffee) in order to select those which better separate food regarding geographical zone or variety. Then, by means of tandem mass spectrometry and accurate mass information, to elucidate a list of selected compounds which allows to create a discrimination statistical model.

- To obtain a wide knowledge about biomarkers related with different nutritional challenges in gilthead sea bream (*Sparus Aurata*), as can be fasting experiments or different feed composition.

- To evaluate possibilities and requirements that the addition of a fourth dimension (Ion Mobility) have in data treatment as well as to the elucidation workflow of selected compounds.

- To obtain indirect biomarkers of illicit drugs consumption (new synthetic cannabinoids) by means of the study of different herbal blends, mainly employed to spice these substances, in smoker's saliva.

Working plan

In order to achieve these objectives, the working plan to perform all these experiments was as follows:

- To study extraction conditions for all the matrices, with the main goal to lose as minimum amount of compounds as possible at this step but covering a wide range of compound polarities.

- To inject samples randomly to avoid differences generated by injection order (systematic errors).

- To study and optimize data pre-processing parameters (peak picking, retention time alignment, normalization,..).

- To evaluate data variance by means of non-biased techniques (Principal Component Analysis, PCA) with the main objective of control the variance in each injection set with "external standard" or QC sample, which is a pool of all the samples composing the

experiment, with the goal to create an average sample, representative of the rest of the set but also employed to control instrumental drift.

- To develop statistical models, univariate (ANOVA) or multivariate (PLS-DA, OPLS-DA, DD-SIMCA), which will depend on the individual objective in each experiment.

- To elucidate as maximum amount of compounds as possible in each experiment.

- To confirm, if possible, the tentative identifications with comparison to commercially available reference standards.

CAPÍTULO I

INTRODUCCIÓN

I.1. Visión general y estado actual

Con el avance de la ciencia y la tecnología, tanto en el ámbito del conocimiento como de la instrumentación disponible, las tecnologías "ómicas" han permitido solucionar problemas que hasta el momento resultaban difíciles y tediosos de abordar. Las "ómicas", que en orden de aparición han sido la genómica, transcriptómica, proteómica y metabolómica, permiten detectar cambios de genes, transcritos, proteínas y metabolitos, respectivamente. De este modo se puede trabajar con una ingente cantidad de datos de una manera no dirigida, obteniendo una visión general en cada estudio. Con el paso de los años, estas ciencias se han ido mejorando e incluso integrando entre ellas, para obtener una visión de conjunto que aporte mucha más información y permita abordar problemas que incluyan compuestos endógenos y exógenos. Este es el caso de las llamadas "foodomics", que combinan las cuatro disciplinas para obtener una visión mucho más global de cómo afecta la nutrición (compuestos exógenos) a un sistema biológico (compuestos endógenos) (Capozzi & Bordoni, 2013).

Para definir el campo de acción de cada una de estas disciplinas, debemos tener en cuenta el tipo de compuestos que se estudian. En este sentido, la genómica se podría definir como el estudio sistemático del conjunto de genes (genoma) de un organismo (Altelaar, Munoz, & Heck, 2012). El objetivo es el análisis del genoma con el objetivo de investigar las interacciones que se producen entre los genes o incluso con el medio ambiente. Sin embargo, no es solo la información contenida en los genes quien gobierna los procesos biológicos, ya que las moléculas de mRNA, que estudia la transcriptómica, y las proteínas, que estudian la proteómica, tienen un papel primordial a la hora de entender comportamientos biológicos más allá del propio ADN (Gygi, Rochon, Franza, & Aebersold, 1999). Finalmente, la llamada "cascada biológica" termina en moléculas de bajo peso molecular (normalmente por debajo de 1500 Da), que varían en función de todos los factores que se relacionan con un determinado estado biológico. Este enorme grupo de compuestos se estudian a través de la metabolómica (Kell, 2004). Sin embargo, así como genómica, transcriptómica y proteómica son tecnologías utilizadas exclusivamente en organismos biológicos, la metabolómica se expande más allá, llegando incluso al estudio de sustancias que nada tienen que ver con procesos biológicos endógenos. En este sentido, podemos encontrar experimentos metabolómicos aplicados al medio ambiente, como pueden cambios en la

contaminación del aire (Vrijheid, 2014). Así pues, y dado el objetivo de esta tesis doctoral, vamos a profundizar en las oportunidades y los requerimientos que tiene, en concreto, la metabolómica.

La metabolómica nace a raíz de la expansión de las otras tecnologías ómicas, cuando el requerimiento de seguir profundizando en los procesos biológicos se hace patente. Así pues, la metabolómica se define inicialmente como el análisis de los metabolitos de un organismo para identificarlos y cuantificarlos (Bino et al., 2004). Sin embargo, con el paso del tiempo la metabolómica se ha bifurcando en: i) en el estudio de perfiles metabolómicos (*metabolomics profiling*) con el objetivo de conocer un elevado número de compuestos predefinidos (*targeted metabolomics*) y así poder conocer la muestra más a fondo y ii) huella dactilar metabolómica (*metabolomics fingerprinting*), donde se pretende obtener un grupo de compuestos que realicen la función de huella dactilar, es decir, que solo con ellos seamos capaces de clasificar o reconocer esa muestra frente a otras diferentes, todo ello de manera no dirigida, *untargeted metabolomics* (Dettmer, Aronov, & Hammock, 2007).

Se ha llegado incluso a clasificar la metabolómica dependiendo de los compuestos analizados, como es el caso de la lipidómica (Q. Shen et al., 2013). Sin embargo, la enorme diversidad de naturalezas químicas de los compuestos (ácidos inorgánicos, lípidos, compuestos volátiles,…) ha requerido el uso de técnicas muy diversas y con un amplio rango lineal, ya que se estima que los metabolitos presentes en una muestra pueden tener entre 7 y 9 órdenes de magnitud de diferencia entre ellos (Dunn & Ellis, 2005). Los avances que las técnicas analíticas han aportado en los últimos años han sido imprescindibles a la hora de desarrollar satisfactoriamente la metabolómica, requiriendo instrumentos con grandes capacidades de adquisición de datos. Igualmente indispensable ha sido el desarrollo de la bioinformática para la minería de datos o las bases de datos *on-line,* así como las herramientas de fragmentación *in-silico*, de las que hablaremos más adelante en la tesis.

I.2. Estructura general de la Tesis Doctoral y campos de aplicación

La presente tesis doctoral se basa en la optimización y utilización de la metabolómica con fines diversos (estudio metabolómico de la nutrición de doradas en cultivo (Capítulo II), autentificación de alimentos (Capítulo III) y obtención de marcadores para el control de nuevas sustancias psicoactivas (Capítulo IV)). Sin embargo, la metabolómica desde su aparición se ha ido empleando cada vez más en estudios, no solo de autentificación de los alimentos (Castro-Puyana, Pérez-Míguez, Montero, & Herrero, 2017; Cevallos-cevallos, Etxeberria, Danyluk, & Rodrick, 2009; Danezis, Tsagkaris, Brusic, & Georgiou, 2016) o nutrición en animales (Chetwynd & David, 2018; Mancano, Mora-Ortiz, & Claus, 2018), sino en muchos otros ámbitos como puede ser la medicina, relacionando compuestos exógenos (tóxicos) con el metabolismo (Bloszies & Fiehn, 2018; Wang et al., 2018) o interacciones entre la nutrición y el metabolismo (Pimentel, Burton, Vergères, & Dupont, 2018).

En esta sección se pretende profundizar en el trabajo realizado anteriormente para comprobar el estado actual de dichos campos de estudio a la par de la necesidad y justificación de estos estudios en los diferentes campos de aplicación en los que esta Tesis Doctoral se ha desarrollado.

I.2.1. Estudio de diferentes estados de nutrición y composición de dietas en doradas

Los trabajos enmarcados en el capítulo II se centran en la obtención de compuestos afectados por diferentes retos metabólicos en doradas, así como su posterior elucidación y explicación biológica.

En este campo de aplicación, junto con otras aproximaciones más convencionales (actividades enzimáticas, evaluación histopatológica, bioquímica de la sangre, etc), es cada vez más habitual el uso de la transcriptomica y de la proteomica para definir el estado nutricional y de salud y bienestar de los peces en cultivo a lo largo del ciclo de producción (Benedito-Palos, Calduch-Giner, Ballester-Lozano, & Pérez-Sánchez, 2013; Calduch-Giner, Sitjà-Bobadilla, & Pérez-Sánchez, 2016; Estensoro et al., 2016; Piazzon et al., 2017; Simó-Mirabet, Perera, Calduch-Giner, Afonso, & Pérez-Sánchez, 2018). Este tipo de aproximaciones también incluye el estudio de los efectos de la dieta sobre la microbiota intestinal (Piazzon et al., 2017), con el objetivo de definir una población de referencia indicativa de un buen estado general del animal. En paralelo, el interés por la aplicación de la metabólica en el campo de la acuicultura ha ido en aumento, y aunque de forma todavía incipiente, el número de publicaciones basadas en Resonancia Magnética Nuclear (NMR) o de Espectrometría de Masas (MS) ha crecido exponencialmente y de forma similar desde la primera publicación del 2002, tal y como se muestra en la **Figura I.1**. Como ejemplos de ello, se han publicando diferentes estudios basados en cómo afecta el ayuno a la trucha arcoiris (Baumgarner & Cooper, 2012) cambios en la dieta, ya sean los efectos que tiene sobre estos diferentes niveles de compuestos en dieta, como por ejemplo en carpa común (Cajka et al., 2013), diferentes fuentes de proteínas en salmón atlántico (M.-B. S. Andersen et al., 2013) o otros micronutrientes como la arginina en salmón atlántico (S. M. Andersen, 2015), la taurina en tilapia del nilo (G. Shen et al., 2018) o incluso acidos grasos, como el araquidónico en pez zebra (Adam, Lie, Moren, & Skjærven, 2017). También se ha estudiado como afecta la introducción xenobióticos en los animales (Huang et al., 2016), como arsénico (Li et al., 2016).

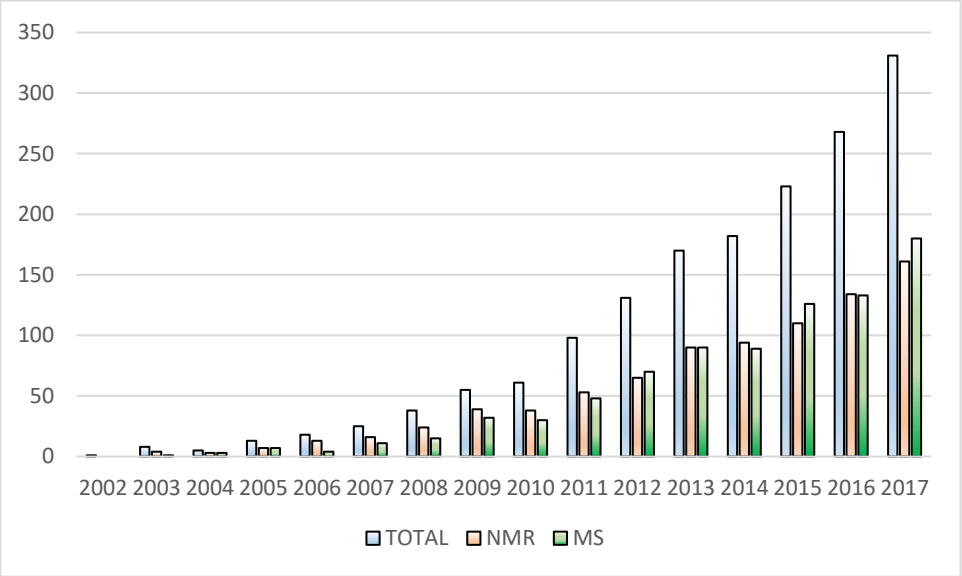


Figura I.1: Resultados de la búsqueda de "*Fish*" y "*Metabolomics*" en *Scopus* (barras azules), y "NMR" (barras naranjas) o MS (barras verdes).

En esta Tesis Doctoral se han estudiado en la dorada los efectos de diferentes desafíos nutricionales sobre los metabolitos presentes en el suero de doradas. Ello ha dado lugar a dos artículos que se enmarcan dentro de un proyecto desarrollado por nuestro grupo de investigación en una estrecha colaboración de la Universitat Jaume I y del Instituto de Acuicultura Torre de la Sal, perteneciente al Consejo Superior de Investigaciones Científicas.

En el artículo científico 1 se aplica la metabolómica no dirigida para identificar los marcadores del suero que se ven más alterados por un ayuno a corto plazo. En el artículo científico 2, se muestran los efectos de la sustitución de harinas y aceites de pescado por ingredientes terrestres de origen vegetal, utilizándose para ello una aproximación no dirigida frente a otra focalizadas en vitaminas de origen tanto dietario como bacteriano. Ello ha puesto de manifiesto que los cambios en respuesta al ayuno son mucho más robustos que los observados en respuesta a las nuevas formulaciones de piensos basados en ingredientes sostenibles que promueven el desarrollo de una acuicultura verde.

I.2.2. Autentificación de alimentos

Los trabajos enmarcados en el capítulo III hacen referencia a la creación de tres modelos estadísticos de autentificación de alimentos por medio de la metabolómica no dirigida. Desde que apareció la metabolómica como una herramienta útil de autentificación, se observa un aumento paulatino de las publicaciones donde ésta es la base para filtrar y obtener marcadores validados. Desde las primeras 4 publicaciones del año 2001 se ha llegado hasta el año 2017, donde se han reportado 1492 publicaciones, como se observa en la **Figura I.2**.

En el año 2007 se desarrolló el primer método mediante metabolómica dirigida para evaluar el origen de los aceites de oliva tunecinos (Brunella Cavaliere et al., 2007). Recientemente, se han publicado algunos artículos evaluando diferentes aspectos que afectan la calidad o el sabor de los aceites de oliva, como puede ser el cruce de variedades (Sánchez de Medina, Riachy, Priego-Capote, & Luque de Castro, 2013) o la determinación de la calidad de éste comparando compuestos fenólicos (estrategia dirigida) (Monasterio, Olmo-García, Bajoub, Fernández-Gutiérrez, & Carrasco-Pancorbo, 2017; Olmo-García, Bajoub, Monasterio, Fernández-Gutiérrez, & Carrasco-Pancorbo, 2017) o por medio de una estrategia no dirigida de sus compuestos volátiles (Garrido-Delgado, Dobao-Prieto, Arce, & Valcárcel, 2015; Sales et al., 2017).

Con el objetivo de desarrollar modelos de autentificación de alimentos se llevaron a cabo tres artículos científicos incluidos en esta Tesis Doctoral. En el artículo científico 3 se aplica la metabolómica no dirigida para obtener marcadores que permitan diferenciar el origen de las diferentes zonas productoras de aceite de oliva de España. En el artículo científico 4 se presenta un set de compuestos que permiten diferenciar almendras españolas de otras extranjeras. Por último, el artículo científico 5 se centra en la obtención de marcadores, en granos verdes de café de Colombia, con el objeto de establecer un modelo analítico fiable que permita asegurar la denominación de origen en un producto de alta calidad como es el café de Colombia.

Los dos primeros artículos se enmarcan dentro de un proyecto desarrollado por nuestro grupo de investigación en la Universidad Jaume I para desarrollar modelos de autentificación de alimentos Premium de la comunidad Valenciana. La zona del interior de Castellón basa una buena parte de su economía en el cultivo de secano, tanto de la aceituna para la producción de aceite de

oliva (Artículo científico 3), como de la almendra (Artículo científico 4); mientras que la parte costera de la provincia es zona productora de naranja (Díaz et al., 2014). Con este objetivo, llevar a cabo un modelo de autentificación para asegurar la variedad de almendras o aceites de oliva, que también afecta a su calidad y precio, proporcionaría una herramienta cuantitativa que aseguraría a las empresas productoras y/o comercializadoras una mayor trazabilidad del producto final. (Galeano Diaz, Durán Merás, Sánchez Casas, & Alexandre Franco, 2005). Han aparecido algunos artículos de clasificación y autentificación de aceites de oliva y almendras dependiendo del campo de cultivo por medio de sus compuestos volátiles por GC-MS (Beltrán Sanahuja, Ramos Santonja, Grané Teruel, Martín Carratalá, & Garrigós Selva, 2011; Galeano Diaz et al., 2005) o con técnicas estadísticas multivariantes donde se analizaron diferentes características físico-químicas (Barreira et al., 2012).

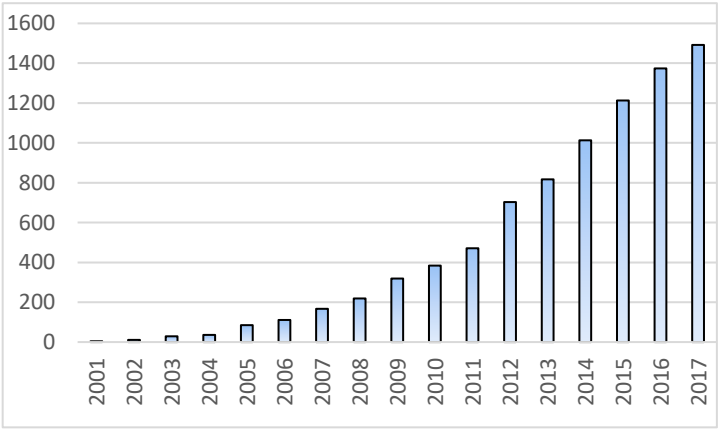


Figura I.2: Resultados de la búsqueda de "*Food*" y "*Metabolomics*" en *Scopus*.

El artículo científico 5 se desarrolló en estrecha colaboración con Cenicafé (principal comercializador de café de Colombia) y la Universidad de Antioquia. El objetivo era obtener marcadores para la discriminación de cafés, para lo cual se habín publicado recientemente en la literatura algunos artículos científicos, en el caso de los cafés de Colombia (Oberthür et al., 2011), de Brasil (Nagai et al, 2016) o incluso para el codiciado café de civeta (Jumhawan et al., 2013). Sin

embargo, en este trabajo se han utilizado más zonas productoras que el artículo de Oberthür del 2011 utilizando un instrumento que permite la elucidación de los marcadores utilizados, aportando un modelo fiable ampliado y mucho más selectivo.

I.2.3. Obtención de biomarcadores en productos con canabinoides sintéticos

Los trabajos incluidos en el capítulo IV se enmarcan en la elaboración y prueba de la capacidad de la metabolómica no dirigida para encontrar marcadores diferenciadores de muestras, no solamente biológicas, sino incluso ambientales. La metabolómica se ha utilizado, en este caso, para encontrar marcadores indirectos de consumo de cannabinoides sintéticos (SCs).

Los cannabinoides sintéticos (SCs) han aparecido en la sociedad actual como sustitutos del cannabis (EMCDDA, 2009), debido a la prohibición del mismo en la mayoría de los países. Sin embargo, estos compuestos, con estructura bastante diferente al THC (compuesto activo en el cannabis, **Figura I.3**.), pero de actividad igual o superior, van surgiendo constantemente en el mercado por la demanda de nuevas drogas que no estén prohibidas. Este problema no surge tanto de la ilegalidad de las mismas sino del requerimiento de determinados sectores (presos, jóvenes en centros de menores, conductores multados por consumo de drogas, …) de poder consumir drogas sin ser detectados por los test a los que son sometidos.

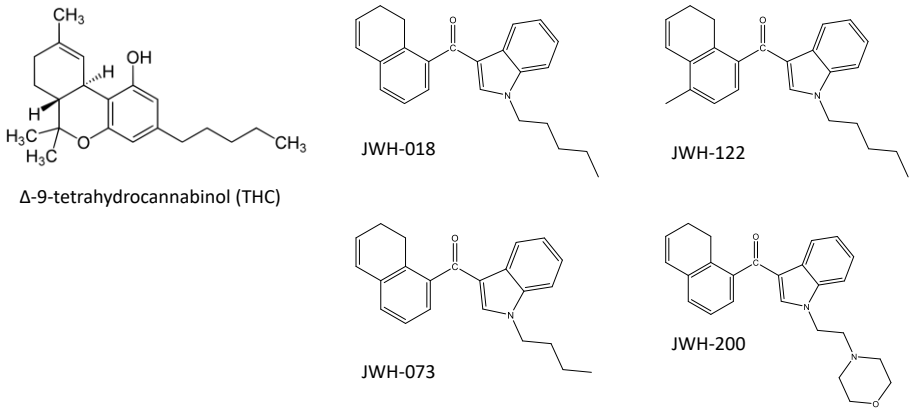


Figura I.3: Estructura química del THC (izquierda) y de cuatro cannabinoides sintéticos (derecha)

Por ello los cannabinoides sintéticos representan una diana en constante movimiento, donde los compuestos a ser controlados se modifican haciendo difícil que la ley actúe sobre éstos (Bijlsma et al., 2017), ya que cuando podrían ser ilegalizados, aparecen otros que no lo están (*legal highs*) y con efectos similares. Teniendo en cuenta que los SCs se venden pulverizando una disolución de los mismos sobre mezclas de hierbas, supuestamente psicotrópicas, y que las hierbas empleadas se reducen a un pequeño grupo de especies, se propone una nueva aproximación para detectar el consumo de SCs desconocidos en base a marcadores de las hierbas base que se utilizan habitualmente. Se establecieron cuatro matrices interesantes sobre las cuales trabajar, el humo producido por las hierbas, el exhalado pulmonar tras expulsar este humo de los pulmones, la saliva de los fumadores y la orina de éstos.

Los artículos de este capítulo se centran en la obtención de marcadores que permitan diferenciar las hierbas utilizadas como base de estos productos (*herbal blend*) del tabaco, ya que el consumo de éste último no está prohibido. Sin embargo, el objetivo es no hacer uso de los canabinoides en sí como marcadores, ya que estos compuestos están en contínuo desarrollo y actualización, mientras que un compuesto procedente de la hierba aparecerá independientemente del compuesto con que se pulverice, en alguna de estas matrices.

En el artículo científico 6 se ha desarrollado el método y se ha demostrado su potencial con compuestos atrapados en la saliva humana. La estrategia aplicada se validó en muestras de saliva de tres voluntarios no fumadores a los que se les tomaba muestra de saliva antes y después de consumir seis de las hierbas utilizadas para vender los SCs. Además, también consumieron tabaco con el objetivo de ser capaces de obtener un marcador de consumo de las hierbas que no apareciera en el tabaco ni en muestras blancas de saliva antes de fumar.

En el artículo científico 7 se propuso seguir con la estrategia en el humo producido por la combustión de estos. En este caso se ha llevado a cabo un estudio con un mayor rango de hierbas (en este caso 14), las cuales también se han reportado como bastante utilizadas (EMCDDA, 2015; Ogata et al., 2013). Además se decidió comenzar por la matriz en la que las diferencias corresponderán puramente a la composición de las hierbas y el tabaco, el humo producido por éstas.

Además de utilizar el concepto expuesto en el artículo científico 6, se valoró la utilización de instrumentos de UHPLC-HRMS con celda de separación por movilidad iónica (D'Atri et al., 2017). Estos instrumentos aportan una cuarta dimensión de separación de los componentes de las muestras, proporcionando no solamente mejor separación y espectros más limpios, sino también una mayor capacidad de elucidación de los mismo en combinación con herramientas de predicción, tanto del tiempo de retención (Bade et al., 2015) como de la sección transversal de colisión (CCS) (Bijlsma, et al., 2017).

I. 3. *Workflow* analítico

En este campo resulta clave definir un flujo de trabajo común, que se debe desarrollar para obtener resultados útiles y concluyentes evitando así errores, que comienza por el diseño experimental y termina por la validación de los resultados. Todo este proceso va a ser expuesto y discutido a lo largo de este capítulo.

I.3.1. Diseño experimental: toma y tratamiento de muestra

El flujo de trabajo de la metabolómica comienza por el diseño experimental. Esta parte del flujo es fundamental y de ella depende en gran medida que el experimento resulte o no satisfactorio. Para comenzar, se debe conocer perfectamente el tipo de muestras con el que se va a trabajar, ya que cuanta más información se tenga sobre ellas, menor probabilidad habrá de cometer errores que comprometan nuestros resultados. En este sentido, debemos asegurar que las muestras tienen unas características comunes (p.e. misma matriz de alimento, no mezclar suero con plasma, …), ya que de lo contrario las diferencias no controladas podrían enmascarar las diferencias en los compuestos de interés. Este rasgo, llamado homogeneidad, permite trabajar con un elevado número de muestras y así obtener resultados más robustos. Sin embargo, en determinados tipos de experimentos, como se verá en el capítulo IV, puede no ser tenido en consideración, y trabajar con grupos heterogéneos, de manera que los marcadores que sean capaces de diferenciar los grupos, si se encuentran, muestren una agrupación más lábil, pero que sea capaz de agrupar muestras *a priori* bastante diferentes entre sí.

A pesar de toda la información que podemos intentar recabar sobre las muestras, una de las limitaciones que tiene la metabolómica no dirigida es que, debido a sus características intrínsecas, se desconoce por completo la naturaleza de los compuestos que serán relevantes al final del proceso. Por este motivo, es importante realizar un tratamiento de muestra universal y sencillo o, de no ser posible, realizar varios tratamientos sencillos, para poder abarcar la mayor amplitud de polaridades posible en lo que a los compuestos se refiere (Raterink, Lindenburg, Vreeken, Ramautar, & Hankemeier, 2014) (**Figura I.4**) .

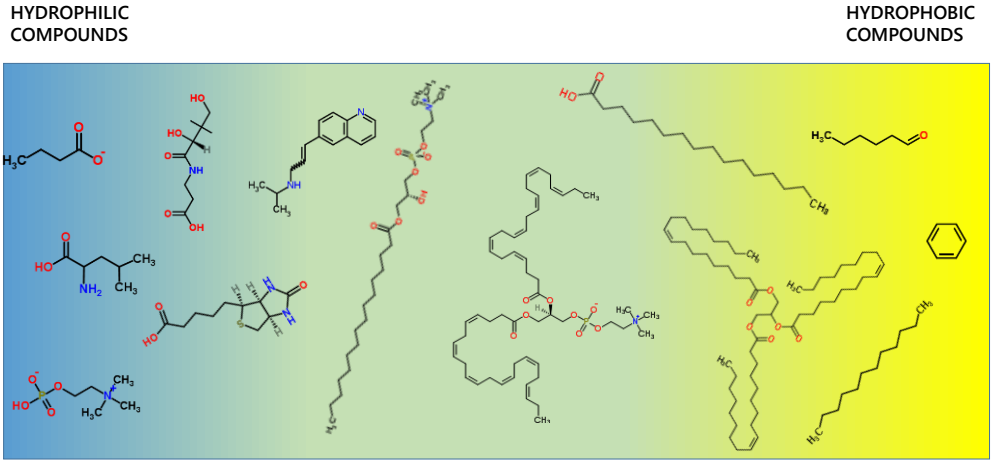


Figura I.4:Rango de polaridad para los compuestos, desde muy polares (izquierda) a completamente apolares (derecha).

En este sentido, se puede hablar de diferentes tipos de matriz y cuál puede resultar el mejor tratamiento de muestra, buscando siempre la sencillez, de los que en la presente Tesis Doctoral se han utilizado:

Alimentos

Probablemente las matrices de alimentos pueden llegar a ser las más complejas de afrontar. La mayor parte de los alimentos son sólidos, por lo que resulta inevitable realizar una extracción y, en caso de alimentos líquidos (aceites, grasas, bebidas,...), la concentración de sus componentes o incluso la incompatibilidad con la inyección directa en las técnicas que se utilizan hacen necesaria una etapa de extracción. Además, debido a sus características, muchos alimentos tienen una parte lipídica importante y una parte polar más o menos homogénea, por lo que resultará difícil realizar un único tratamiento de muestra. Por ello, afrontar este paso con dos extracciones diferentes parece ser una solución adecuada (Cevallos-cevallos et al., 2009), ya que permitirá analizar un amplio rango de polaridades de forma separada, teniendo siempre en cuenta

que es inevitable dejar compuestos fuera del rango de análisis a la vez que lo será tener algunos analitos en ambas fracciones.

En el caso de los alimentos sólidos, una extracción para los compuestos polares (zona izquierda de la **Figura I.4**) deberá contener una parte importante de disolvente polar como agua, metanol, etanol o acetonitrilo. Se puede acidificar el disolvente con el fin de protonar los compuestos y que estos aumenten su carácter hidrófilo. De este modo los compuestos que se extraerán en el procedimiento tendrán un carácter más bien polar dentro de un rango de polaridades dilatado, donde se encuentran compuestos cargados y/o pequeñas moléculas polares o semi-polares.

Por otro lado, se puede realizar una extracción con disolventes más apolares, como la acetona, el butanol o el cloroformo, para analizar las familias de compuestos que no se extraen correctamente con disolventes polares. Esta segunda extracción nos permitirá abarcar un rango de polaridades en una zona más bien apolar (zona derecha de la **Figura I.4**), en la que analizaremos ácidos grasos, triglicéridos o incluso compuestos volátiles.

Si fuese necesario, para aislar compuestos presentes en los alimentos en concentraciones muy bajas, como el caso de los micronutrientes (vitaminas…), se podrían utilizar técnicas de extracción como el SPE (*Solid Phase Extraction*). En el campo de la metabolómica no dirigida, el uso de esta técnica puede resultar complicada de justificar. Esto es debido a que la pérdida de analitos puede ser importante a causa de las características de la técnica, que es aislar un determinado grupo de compuestos por sus características químicas. Sin embargo puede resultar necesaria en el supuesto de querer realizar una aproximación dirigida tras la no dirigida, en la que los compuestos de interés no son analizables con una extracción genérica.

En el caso de las muestras líquidas, el procedimiento puede incluso llegar a ser un análisis directo si la matriz no es excesivamente compleja. En caso contrario, como en aceites, el tratamiento ha de ser similar a las muestras sólidas, dependiendo de sus características, con una extracción de los compuestos polares (extracción líquido-líquido), y una dilución de la matriz en disolventes apolares, que permitirán analizar los compuestos de elevada concentración, como en el caso de los aceites, los triglicéridos, diglicéridos, etc. En este sentido, debido a las características

de la técnica de separación utilizada *a posteriori*, se deberá tener en cuenta la compatibilidad del disolvente con la misma, a fin de evitar tener que realizar pasos extra con la consecuente posibilidad de pérdida de analitos.

Biofluidos

Las matrices biológicas pueden resultar más fáciles de tratar, ya que se componen en un gran porcentaje de agua. Este hecho hace que la gran mayoría de sus componentes sean compatibles con la cromatografía de líquidos, a la vez que una simple dilución de la muestra sea suficiente para la inyección (*dilute-and-shoot*). Además, dado el objetivo inicial de la metabolómica, existen una gran cantidad de estudios realizados en matrices biológicas (sangre, orina, saliva,..) y su tratamiento de muestra se ha estandarizado a lo largo del tiempo (Duarte, Rocha, & Gil, 2013).

En este sentido, las matrices como la sangre (suero o plasma) se tratan para eliminar las macromoléculas presentes (ADN y proteínas) con acetonitrilo (ACN), metanol o etanol, evitando así posibles interferencias en el análisis metabolómico, que se centra en moléculas de menor peso molecular (Bruce et al., 2009). La matriz, una vez desproteinizada y centrifugada se puede analizar directamente por cromatografía de líquidos o realizar tratamientos extra para adecuarla a otras técnicas de separación.

En matriz de orina, al no presentar un contenido de macromoléculas tan elevado como el suero o el plasma, se suele realizar una centrifugación para eliminar partículas en suspensión y una dilución que depende de la técnica que utilizaremos (Ramses F. J. Kemperman et al., 2006).

Otras matrices biológicas, como la saliva, en las que el contenido de agua es muy cercano al 100%, es necesario un tratamiento de muestra que elimine macromoléculas presentes en la matriz, a la vez que concentre el resto de los compuestos. En este sentido, las matrices de saliva se desproteinizan con ACN y el sobrenadante se lleva casi a sequedad, reconstituyendo a la vez que preconcentrando con un disolvente compatible con la técnica que se va a utilizar (Malkar et al., 2013). Este procedimiento, al ser más agresivo con la muestra, puede conllevar pérdida de compuestos. Sin embargo, la concentración a la que se encuentran la mayoría de los compuestos presentes en ésta hace muy complejo trabajar sin preconcentrar la muestra.

Muestras ambientales (gases)

En este caso, al no tratarse de una muestra tan fácil de manejar como los sólidos o los líquidos, los procedimientos se basan principalmente en las extracciones gas-sólido. Las matrices gaseosas se hacen circular a través de un sólido adsorbente (SPE, SPME, SBSE,…) y se eluyen posteriormente en un disolvente adecuado a los requerimientos de la técnica para el caso de la cromatografía de líquidos, o incluso la desorción térmica de fibras o de sólidos en los que previamente se han atrapado los compuestos, si se está trabajando en cromatografía de gases. .

Generación de la muestra QC

En este momento también se genera una muestra adicional, llamada QC. Esta muestra, que es simplemente una mezcla del resto de muestras o extractos (una muestra representativa del grupo), es necesaria, como se verá posteriormente, para estabilizar la columna en las primeras inyecciones, ya que hay experimentos donde la cantidad de muestra es muy limitada (unos pocos microlitros). Al crear este QC se puede obtener una buena cantidad de muestra (si utilizamos 10 µL de cada muestra y nuestro batch se compone de 50 muestras por ejemplo, generamos 0.5 mL de muestra para estabilizar la columna. Además de ello, también se utiliza para controlar la correcta normalización de los datos (Díaz et al., 2016; Díaz, Pozo, Sancho, & Hernández, 2014).

I.3.2. Técnicas analíticas

Debido a la ingente cantidad de compuestos químicos que existen y a su diversidad en cuanto a características físico-químicas, no se puede seleccionar una técnica universal con la que abordar cualquier problema en metabolómica. Por ello, se han utilizado una amplia variedad de técnicas entre las que se encuentra la resonancia magnética nuclear (RMN), la espectrometría de masas (MS) acoplada a técnicas de separación, como la cromatografía de líquidos (LC), la cromatografía de gases (GC) o la electroforesis capilar (CE), o incluso la espectroscopia de infrarrojos (IR) (Dunn & Ellis, 2005). Todas estas técnicas pueden, combinadas, aportar una detección universal a la metabolómica. Sin embargo, cada una de ellas por separado aporta unas características específicas que deben tenerse en cuenta a la hora de seleccionar una o varias de ellas.

En esta Tesis Doctoral, la experiencia en espectrometría de masas del grupo de investigación unida a la mayor sensibilidad que ésta presenta frente a RMN, ha hecho que la técnica de elección haya sido MS, a pesar de que el empleo de varias técnicas puede proporcionar información adicional valiosa. Además, la mayor afinidad de las matrices biológicas (sangre, orina, saliva) y el amplio rango de polaridades que la cromatografía de líquidos puede abarcar frente a GC o CE ha hecho que los esfuerzos se hayan centrado en el uso de esta técnica en concreto. Sin embargo, en otras líneas de investigación dentro del grupo se está trabajando en metabolómica con cromatografía de gases, poniendo de manifiesto la necesidad que combinar técnicas analíticas.

En este sentido, la cromatografía de líquidos y la espectrometría de masas son las técnicas que se van a abordar más en detalle en las siguientes páginas.

Cromatografía de líquidos

A pesar de la alta selectividad que proporciona la espectrometría de masas de alta resolución, como se verá en este mismo capítulo, se hace necesario el uso de una técnica de separación si tratamos con matrices complejas, que permita separar los compuestos de una muestra antes de ser medidos/detectados por el espectrómetro de masas. En este sentido, la cromatografía surge como una herramienta íntimamente ligada al sistema de detección instrumental.

Esta técnica de separación basa su funcionamiento en la interacción de los compuestos a separar, con grupos químicos presentes dentro de las columnas de cromatografía (fase estacionaria) a la vez que con el líquido (fase móvil) que circula por su interior (Ardrey, 2003) . De este modo, los compuestos que se pueden separar por LC deberán presentarse en disolución, por lo que podrán ser incluso compuestos con cierta polaridad y baja volatilidad, además de termolábiles. Los compuestos volátiles, termoestables y apolares preferentemente son abordados por cromatografía de gases.

La fase móvil en la cromatografía de líquidos, al contrario que en GC, juega un rol protagonista en la separación, ya que es la encargada de "arrastrar" los compuestos a lo largo de

la columna mediante su interacción con los mismos. Esta característica, llamada fuerza eluotrópica, dependerá de la polaridad de la fase móvil. Estas polaridades, como se muestra en la **Tabla I.1**, varían bastante entre los disolventes y será la combinación de ellos la que hará que los compuestos eluyan de la columna cromatográfica en el orden deseado. Para ello se podrá aplicar una elución isocrática, donde las proporciones de fases móviles no varíen a lo largo de la separación, o un gradiente donde las proporciones si varíen, de forma que podremos modular el paso de estos a lo largo de la columna dependiendo de las necesidades.

Índice de polaridad	Disolvente	Miscibilidad con agua (%m/m)	Viscosidad (cP)	Punto de ebullición ºC
9.0	Agua	--	1.00	100
6.6	Metanol	100	0.60	64.7
6.2	Acetonitrilo	100	0.37	81.6
3.9	Butanol	0.43	3.01	177.7
0.0	n-hexano	0.001	0.313	68.7

Tabla I.1: Características físicas de algunos disolventes utilizados en LC

Sin embargo, la velocidad de avance de los compuestos no dependerá únicamente de la fase móvil, sino también de la fase estacionaria, que creará a lo largo de la columna una serie de interacciones adsorción/desorción que serán las encargadas de determinar el tiempo que el analito permanecerá en el interior de la columna.

La fase estacionaria en la cromatografía de líquidos es determinante a la hora de seleccionar las fases móviles y, por ende, afecta al tiempo total del experimento. En este sentido, a pesar de que existen más tipos de mecanismos de separación, se han utilizado principalmente dos en esta tesis.

El primero de ellos es la **cromatografía de fase reversa (RP)**. Este tipo de cromatografía se basa en la utilización de partículas minúsculas (unos pocos µm), de sílice o poliméricas, en las que se unen a los grupos funcionales que hay en la superficie de éstas, pequeñas cadenas alquílicas (C8, C18,…) que crearán una interacción con los grupos apolares de los compuestos a estudiar, como puede observarse en la **Figura I.5**. Cada tipo de columna tiene sus características en cuanto a rango de polaridades que es capaz de separar, rango de pH de

trabajo (ya que las columnas de sílice no pueden trabajar a pH alcalinos pues se disolverían), etc. Todas estas características se deberán de tener en cuenta para la realización de los experimentos.

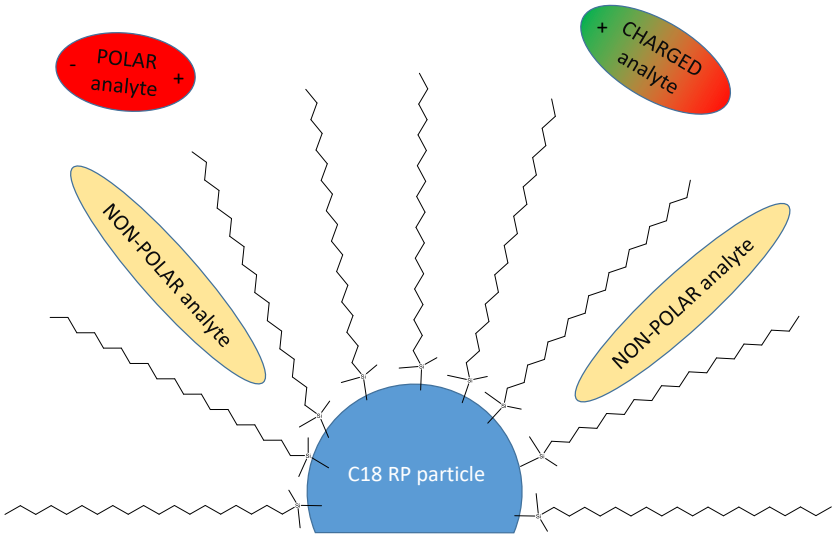


Figura I.5: Partícula de RP (C18), con las posibles interacciones con los compuestos.

Con la columna seleccionada, se consigue que los analitos de interés se queden retenidos en la superficie de las partículas de la fase estacionaria para ser eluidos a lo largo de la columna por la fase móvil. En este sentido, la fase móvil comenzará con un disolvente más polar (como suele ser típicamente agua) en gran porcentaje, al que llamaremos disolvente *A*, para arrastrar fuera de la columna compuestos que no se retienen y que así no interfieran en la detección de los compuestos de interés, y un disolvente *B*, menos polar que el agua (acetonitrilo, metanol, etanol,…), el cual debe ser miscible con *A* y que sea a su vez capaz de interaccionar con los compuestos atrapados en la columna, haciéndolos eluir para ser detectados.

Llegados a este punto, existen dos formas de trabajo en cromatografía de líquidos, pudiendo ser con elución isocrática o con gradiente de concentraciones. En la elución isocrática, los porcentajes de *A* y *B* se mantienen constantes, siendo el tiempo y el porcentaje de B los encargados de arrastrar fuera de la columna los compuestos. Este tipo de modo de trabajo resulta

útil si los compuestos de interés tienen unas características muy parejas, de modo que una pequeña variación de *A* o *B* hará que salgan juntos de la columna. Sin embargo, los compuestos que permanecen más tiempo en la columna presentan menor resolución cromatográfica. En metabolómica, como *a priori* se desconoce la naturaleza de los compuestos, resulta ilógico adoptar este modo de trabajo.

El modo de gradiente de concentraciones, por su parte, implica que el porcentaje de *A* se irá reduciendo a lo largo del experimento, aumentando así el de disolvente *B*, siendo estas variaciones de polaridad global de la fase móvil las que arrastren los compuestos fuera de la columna más rápidamente. Estas variaciones de polaridad en metabolómica suelen ser desconocidas, por lo que un gradiente lineal entre el 10 y el 90% de *B* puede ser una estrategia adecuada. Sin embargo, dependiendo del rango de polaridades de los compuestos a estudiar, es posible utilizar disolventes más apolares que sean capaces de separar compuestos menos polares, los cuales no se eluyen de la columna con metanol o acetonitrilo. En este caso, se puede utilizar metanol como disolvente *A*, para eliminar así "interferentes" de la matriz, que en este caso serían los compuestos analizados en el gradiente expuesto anteriormente, y utilizar como disolvente *B* otros con menor polaridad (2-propanol, butanol, tetrahidrofurano…) que sean capaces de separar compuestos en otro rango de polaridad más bajo, extendiendo así nuestra capacidad de análisis.

Además de la composición de la fase móvil, resulta importante seleccionar modificadores que ayudarán a la hora de favorecer la aparición de especies pseudomoleculares, como el ion protonado ($[M+H]^+$) o desprotonado ($[M-H]^-$), como pueden ser ácidos orgánicos (ácido fórmico, acético…) o bases orgánicas (amoniaco, …) en concentraciones muy pequeñas (0.01 %) (Poole, 2003). Estos deben ser volátiles para no ensuciar la interfase, a la vez que ayudan a controlar el pH y pueden dar/aceptar protones para evitar la formación de aductos con sodio ($[M+Na]^+$) o potasio ($[M+K]^+$), que presentan espectros de fragmentación pobres.

Por otro lado, se encuentra la **cromatografía líquida de interacción hidrofílica (HILIC)**, en la que el modo de trabajo es opuesto a RP. Este tipo de cromatografía, que ha aparecido como alternativa a la cromatografía de fase normal, tiene partículas de sílice que pueden estar sustituidas por grupos amino, ciano, o simplemente trabajar con los grupos silanol

de la superficie del sílice. Al contrario que RP, tiene su rango de separación en compuestos muy polares, como iones o pequeñas moléculas, típicamente analizables por electroforesis capilar (Buszewski et al, 2012). Además, ya que la cromatografía líquida de fase normal es incompatible con el agua, en HILIC es requisito trabajar con ésta como fase móvil, haciendo posible la separación de compuestos polares en matrices biológicas, que presentan típicamente elevados contenidos de agua. En este sentido, las partículas de fase estacionaria en el mecanismo HILIC se rodean por una pequeña capa de agua, que hace posible varios tipos de interacción, con compuestos iónicos, compuestos neutros, con los que interacciona por enlaces de puente de hidrógeno (OH del silanol) o compuestos que se adsorben en el agua, como se puede apreciar en la **Figura I.6**.

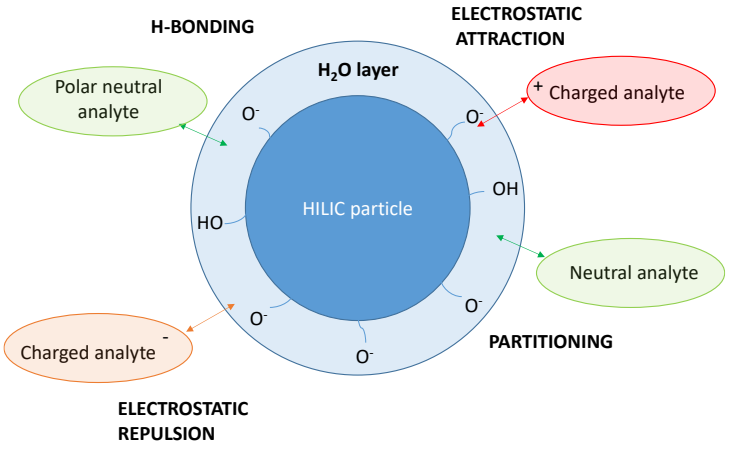


Figura I.6: Partícula de HILIC, con las posibles interacciones con los compuestos.

Por todo esto, la cromatografía HILIC aporta una separación intermedia entre la cromatografía de fase normal y la electroforesis capilar, proporcionando un abanico de polaridades complementario a RP, que hace que ambas trabajen complementándose la una con la otra lo que resulta extremadamente útil.

Una vez creada la capa de agua, se consigue que los analitos de interés se queden retenidos en la superficie de las partículas de la fase estacionaria para ser eluidos a lo largo de la

columna por la fase móvil. En este caso, la fase móvil *A* será un disolvente más apolar que el agua (acetonitrilo, metanol,…) en gran porcentaje, siempre con un mínimo del 5% de agua para mantener la capa de agua en las partículas y no desestabilizar la columna, pero que elimine compuestos que no se retengan en la capa de agua. El porcentaje de *A* se irá reduciendo a lo largo del experimento, aumentando así el de disolvente *B* (agua), haciendo eluir los compuestos polares para ser detectados.

Además de la composición de la fase móvil, resulta imprescindible utilizar modificadores que mantengan la capa de agua estable alrededor de las partículas, como puede ser formiato de amonio (NH_4HCOO) o acetato de amonio (NH_4CH_3COO), en concentraciones elevadas (5-10 mM), que además ayudarán a crear aductos con amonio ($[M+NH_4]^+$) o con formiato ($[M+HCOO]^-$) o acetato ($[M+CH_3\text{-}COOH]^-$). También resultan útiles los buffers con sus ácidos orgánicos conjugados ($HCOOH/HCOO^-$) para controlar el pH, ya que las columnas HILIC son extremadamente sensibles a pequeñas variaciones en pH o concentración de modificadores, haciendo incluso que el orden de elución de los compuestos varíe. Sin embargo, estas elevadas concentraciones, a pesar de ser imprescindibles para la cromatografía, pueden suponer un inconveniente debido a la bajada de señal causada por la deposición de especies en el cono de extracción, haciendo de HILIC una técnica menos robusta y con mayores dificultades técnicas que la RPLC.

Por otro lado, no solo es importante el tipo de partícula que se utiliza, sino también su eficiencia. En este sentido, el parámetro que se utiliza es el número de platos teóricos de la columna. Este factor está ligado a la eficiencia de la columna y se refiere al número de interacciones que un analito puede efectuar a su paso por esta. Un mayor número de interacciones implica que dos compuestos similares puedan o no ser resueltos a la salida de esta, obteniendo mejores separaciones cuanto mayor número de platos teóricos tenga la columna. Este número se puede optimizar:

- *Aumentando la longitud de la columna*. Por otro lado, esto implica aumentar los tiempos de análisis, ensanchando aún más los picos debido a efectos de difusión longitudinal.

- *Disminuyendo el tamaño de partícula*. Partículas menores proporcionan un mejor empaquetamiento, que no solo afecta a la reproducibilidad de las inyecciones, sino también a aumentar los platos teóricos y a disminuir efectos indeseados, como la difusión de Eddy.

Este segundo parámetro ha permitido diseñar columnas mucho más eficientes disminuyendo el tamaño de partícula de 10 µm (LC) a 5 µm (HPLC) e incluso a 1.7 µm (UHPLC), disminuyendo así el volumen muerto de la columna y aumentando el número de interacciones. El único problema de estas columnas es la sobrepresión que aportan, debiendo trabajar con bombas capaces de soportar contrapresiones de incluso 1000 bares.

Sin embargo, superando las dificultades técnicas de la presión y como se puede observar en la **Figura I.7**, el rango de velocidad de las fases móviles para obtener una altura de plato teórico (HEPT) menor aumenta considerablemente, así como la velocidad lineal absoluta, lo que proporciona análisis mucho más eficaces y cortos.

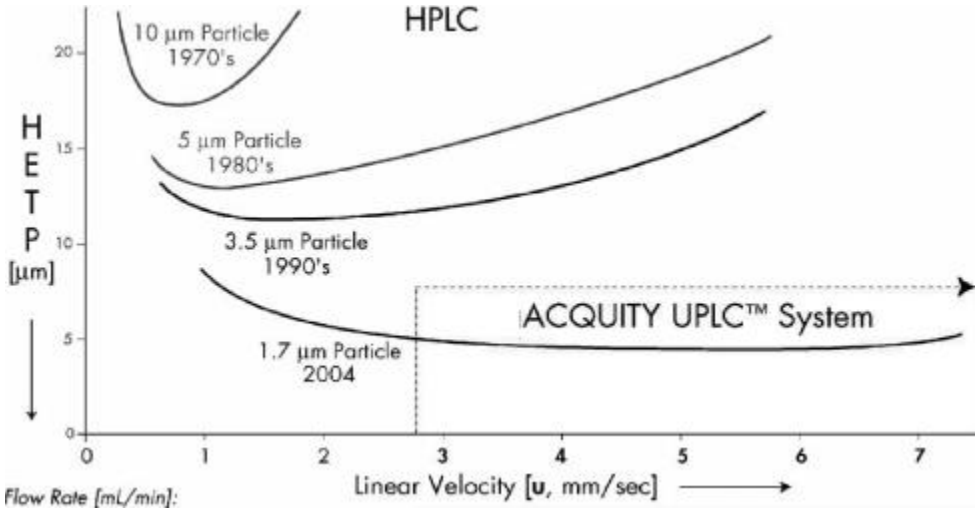


Figura I.7: Gráfica de HEPT frente a velocidad lineal de la fase móvil.

Además de reducir el tamaño de partícula, también es posible reducir el coeficiente de transferencia de masa en la ecuación de Van Deempter para aumentar así la eficiencia, reduciendo el número de poros en las partículas. Por este motivo aparecieron las columnas llamadas *Fused-core* (Kamour, Ammar, El-Attug, & Almog, 2013), que presentan el centro de las partículas sin poros, haciendo que los analitos solo puedan interaccionar en la parte más superficial de la columna (0.5 µm, **Figura I.8**). De este modo, es posible obtener una HEPT menor con tamaños de partícula mayores, típicamente ligados a HPLC (2.7 µm) lo que disminuye la presión del sistema sin perder eficacia.

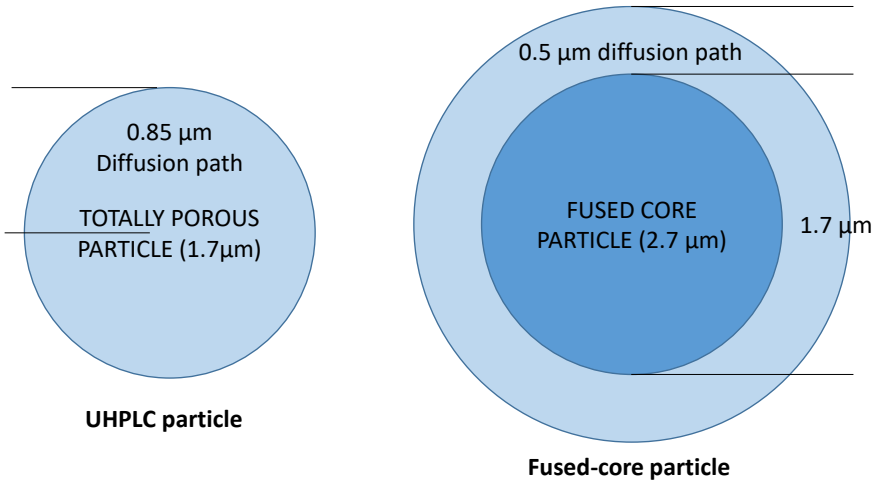


Figura I.8: Comparación de partícula UHPLC (1.7 µm) con fused-core (2.7 µm).

Utilizando la tecnología de UHPLC y/o *Fused-core*, unido a la cromatografía RP y HILIC, ha sido posible analizar un amplio rango de polaridades en matrices tanto biológicas, de alimentos o ambientales obteniendo un enorme poder de resolución y proporcionando una gran cantidad de analitos resueltos que serán detectados posteriormente mediante espectrometría de masas.

Interfases

La espectrometría de masas determina las relaciones masa/carga de los compuestos ionizados (*m/z*) en fase gas, por lo que al interior del instrumento debemos introducir iones en fase gas. Puesto que la cromatografía analiza compuestos que físicamente se encuentran

generalmente neutros, bien sea en fase gas para GC, o bien disueltos en un disolvente en LC, debemos introducir entre ambos instrumentos una interfase encargada de transformar moléculas neutras en iones en fase gas. Para ello, en cromatografía de gases típicamente se han utilizado fuentes de ionización que bombardean los compuestos que salen de la columna en fase gas, con electrones de alta energía (70eV), para conseguir cargarlos positivamente por pérdida de un electrón ($M^{+\cdot}$).

Sin embargo, la cromatografía de líquidos debe conseguir extraer iones de un líquido que fluye constantemente desde la columna a un flujo elevado (típicamente entre 0.3 y 1 mL/min para UHPLC). Para ello, las tecnologías más ampliamente utilizadas en la actualidad son las interfaces a presión atmosférica (API), tanto Electrospray como Ionización Química a Presión Atmosférica (APCI), siendo el electrospray, la interfase que se ha utilizado en los trabajos de esta tesis.

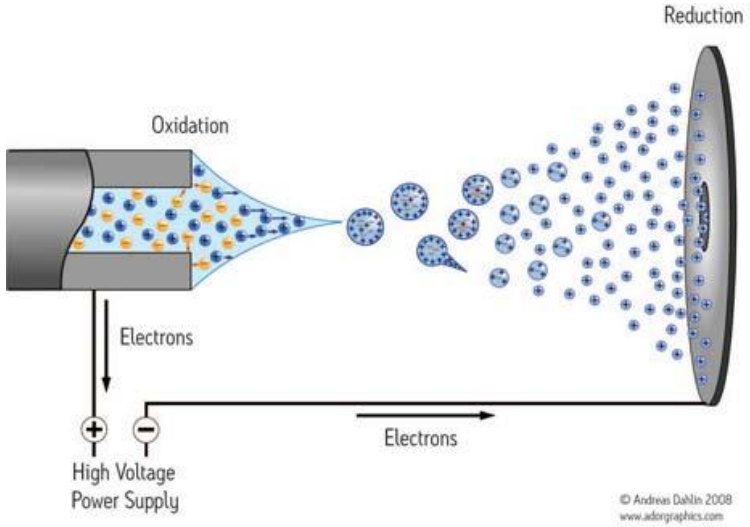


Figura I.9: Representación esquemática del funcionamiento de la interfaz ESI

Esta interfase electrospray (Whitehouse, Dreyer, Yamashita, & Fenn, 1985) consiste en un capilar altamente cargado (entre 0.5 y 3.5 kV típicamente) enfrentado a un cono de extracción que se sitúa dentro de la fuente. Este, unido a un flujo de gas a temperatura elevada (unos 800L/h de nitrógeno a 550 ºC) es capaz de nebulizar el líquido que sale continuamente de la columna

cromatográfica con los iones de una misma carga típicamente (des)protonados o formando aductos con contraiones que se encuentran en la fase móvil, como NH_4^+, Na^+ , K^+ en modo de ionización positivo o $HCOO^-$, CH_3-COO^-, Cl^- en modo de ionización negativo, creando gotas cargadas. Mediante la temperatura aportada por el gas a un elevado flujo, se reduce casi instantáneamente el tamaño de las gotas generadas hasta el punto que las repulsiones entre los iones hacen que la gota estalle, generando iones en fase gas. Estos iones son atraídos por el cono de extracción que los introduce dentro del analizador de masas (**Figura I.9**).

Analizadores de masas

Como ya se ha comentado en este capítulo, los analizadores de masas (MS) permiten el análisis de la relación que existe entre la masa y la carga de un ion (*m/z*). Esta característica hace que, de manera intrínseca, la espectrometría de masas solo sea capaz de analizar compuestos previamente ionizados. Sin embargo, entre los analizadores de masas que tenemos disponibles hoy en día, debemos hacer una distinción, que resulta fundamental, a la hora de seleccionarlos para trabajar en metabolómica. Diferenciamos entre analizadores de baja o de alta resolución, en función de su capacidad para separar entre iones con una relación masa/carga muy similar.

En metabolómica no dirigida tenemos tres requisitos básicos a la hora de escoger el analizador de masas: capacidad de determinar un amplio rango de *m/z,* elevada sensibilidad y selectividad e información estructural aportada. La importancia de ellas resulta clave a la hora de afrontar el momento de la elucidación del/de los compuesto/s seleccionado/s, ya que se desconoce cuál será la masa de éste, la concentración y la coelución con otros compuestos de masa similar. Por este motivo resulta casi imprescindible en metabolómica trabajar con cromatografía de líquidos acoplada con analizadores de alta resolución, mientras que en cromatografía de gases acoplada a espectrometría de masas de baja resolución se puede hacer uso de librerías de compuestos que, si bien ayudan proporcionando la identidad de los compuestos, no se encuentran disponibles en LC-MS.

En este sentido, podemos encontrar varios tipos de analizadores de masas de alta resolución, entre los que se encuentran el tiempo de vuelo (TOF), el Orbitrap o el *Fourier Transform Ion Cyclotron Resonance (FT-ICR)* (Dass, 2007). De entre ellos probablemente el último, a

pesar de su elevadísimo poder de resolución, tiene un coste del mantenimiento muy elevado, lo hace muy difícil su uso en la mayoría de laboratorios. Sin embargo, tanto el TOF como el Orbitrap se han utilizado en metabolómica en los últimos años, llevándose incluso a cabo su comparación en este campo (Díaz et al., 2016; Raro et al., 2015).

Además de disponer de analizadores de alta resolución, con los que podemos obtener un amplio espectro de masas con una resolución muy elevada, es igualmente posible acoplar dos analizadores distintos en serie (uno de baja y otro de alta resolución), dependiendo del objetivo del experimento. En este caso, hablamos de los espectrómetros de masas híbridos. Este tipo de analizadores combinan un analizador capaz de preseleccionar masas, de manera que aplicamos el modo de trabajo de barrido de iones producto, pero ganando en selectividad para el ion preseleccionado con información de masa exacta de sus fragmentos. Esto proporciona información imprescindible a la hora de realizar la elucidación estructural. Entre estos se encuentran el *Q-TOF* o el *Q-Orbitrap,* entre otros.

En esta Tesis Doctoral, debido a que únicamente se ha utilizado como analizador de masas el *Q-TOF,* se va a profundizar en la forma de analizar iones de este último.

Cuadrupolo, TOF y Q-TOF

El analizador de masas de tiempo de vuelo, como ya se ha comentado, proporciona información de masa exacta mientras que el cuadrupolo simplemente es capaz de medir iones con masa nominal. La unión de ambos, sin embargo, es capaz de combinar las características de cada uno en beneficio del usuario final.

El *cuadrupolo* está compuesto por cuatro barras metálicas cilíndricas o hiperbólicas dispuestas en paralelo, a las que se aplica, dos a dos, una corriente continua y un voltaje de radiofrecuencia. Con la aplicación de estos potenciales, los iones que entran por una parte del cuadrupolo se sienten atraídos de manera intermitente por unas u otras barras, de modo que solamente una relación m/z es capaz de atravesarlo sin colisionar. Los iones que atraviesan el cuadrupolo describiendo una trayectoria helicoidal sin colisionar (iones resonantes) y finalmente saldrán del cuadrupolo, mientras que si no cumplen la relación *m/z* colisionan antes de

conseguirlo (iones no resonantes). Este comportamiento se puede observar en la **Figura I.10**. En la parte posterior del cuadrupolo, solamente los iones seleccionados habrán podido seguir una trayectoria estable. Sin embargo, debido a las características del filtro cuadrupolar, su poder de resolución es de 1 Da, e iones con menos de 1 u.m.a. de diferencia podrán conseguir atravesarlo sin colisionar y no podrán ser determinados por separado.

De este modo, el cuadrupolo permite realizar barridos de masas relativamente rápidos, aunque no de manera simultánea, sino secuencial (modo *Full scan*) o seleccionar una relación *m/z* y dejar únicamente pasar iones con una determinada *m/z* (*Selected Ion Monitoring)*.

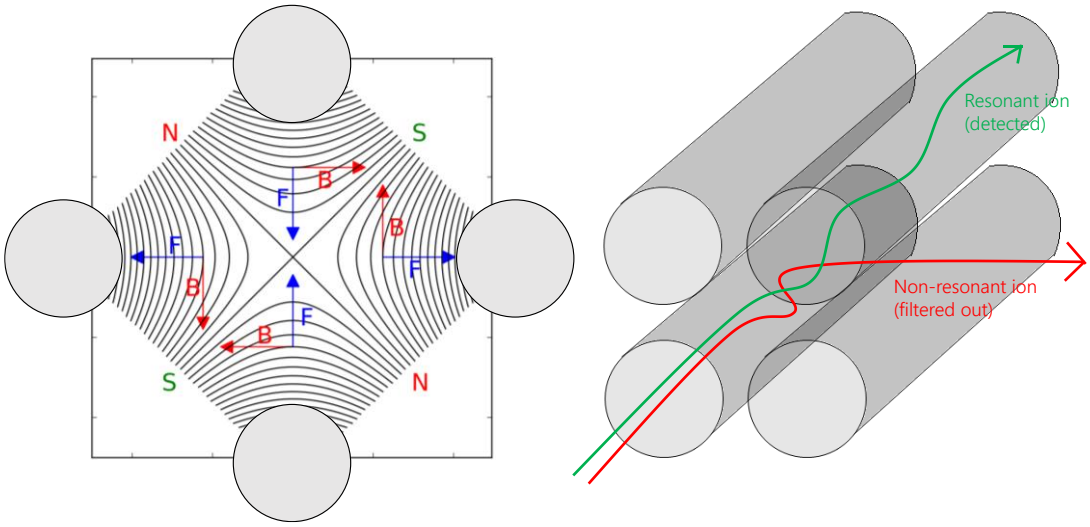


Figura I.10: Analizador cuadrupolar visto desde la entrada de iones (izquierda), donde se observan las fuerzas eléctricas y magnéticas que afectan al ion o en perspectiva (derecha) donde se observa la trayectoria de dos iones, uno resonante y otro no resonante.

El tiempo de vuelo (*TOF*), por su parte, impulsa los iones que entran al analizador aplicando ortogonalmente un potencial eléctrico y acelerándolos. De este modo, aplicando una energía igual a todos los iones, que se transformará en energía cinética, aquellos que tengan menor masa alcanzán mayor velocidad y llegarán antes al detector, mostrando un menor tiempo de vuelo (**Figura I.11**). También existe la posibilidad de añadir un reflectrón que refleje los iones

en sentido opuesto al que entran, describiendo una trayectoria en forma de *V* (como se observa en la **Figura I.11**) o el uso de dos reflectrones para generar una *W*. Con ellos se logra aumentar la longitud de la trayectoria de los iones sin necesidad de utilizar tubos de vuelo más largo, teniendo un equipo compacto con mayores resoluciones. Debido a la precisión actual en el control del tiempo, se puede obtener información del detector en periodos de tiempo muy cortos, lo que hace que su poder de resolución se quede por debajo de 10^{-3} Da (0.3 mDa aproximadamente) (Cotter, 1998). En este tipo de analizadores, se mide el tiempo que un ion tarda en cruzar el tubo de vuelo y hay que esperar, que tanto los iones pequeños como los más grandes, lleguen al detector, haciendo que el equipo solo sea capaz de trabajar en modo de barrido completo (*Full scan*). Además, debido a sus características, el TOF presenta una mejor resolución para iones con una relación *m/z* mayor, siendo ampliamente utilizado para analizar proteínas, pero igualmente válido para moléculas pequeñas (MW<1000Da).

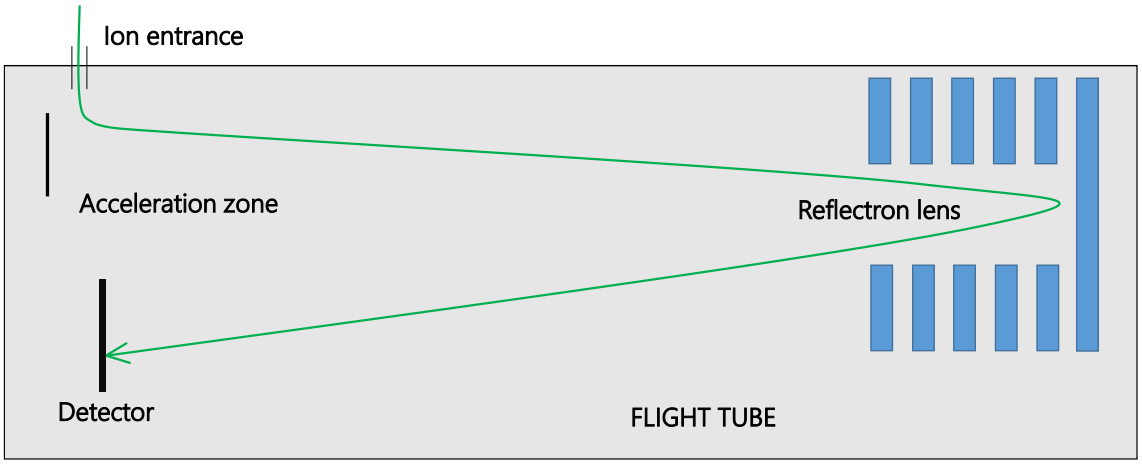


Figura I.11: Esquema básico del analizador de tiempo de vuelo.

Los analizadores híbridos (como el *Q-TOF*), por su parte, pueden trabajar de varios modos. Estos equipos presentan, además del primer (cuadrupolo) y segundo analizador (tiempo de vuelo), otro "analizador" intermedio (inicialmente un cuadrupolo, pero que actualmente puede ser un hexapolo u otros diseños) que actúa como celda de colisión. Así pues, los iones que atraviesan el

primer cuadrupolo chocan contra un gas inerte presente en esta celda. Dependiendo de la energía de colisión que apliquemos estos iones pueden o no fragmentarse dentro de esta celda de colisión. Finalmente, llegan al TOF, donde se obtiene un espectro completo de alta resolución de todos los iones. Existen dos modos de trabajo de estos instrumentos que resultan extremadamente útiles en metabolómica.

El primero, que proporciona una visión muy general de los compuestos presentes en las muestras, es el modo *Full scan*. En este modo de trabajo, los iones pasan a través del cuadrupolo que actúa solo como guía de transmisión de iones, de manera que todos los iones cruzan el cuadrupolo hacia la celda de colisión. Es importante que no perdemos sensibilidad con la actuación del primer cuadrupolo.

Posteriormente, en la celda de colisión, que se encuentra llena del gas inerte, los iones se aceleran con una determinada energía, haciendo que estos se fragmenten o no. Es esta característica la que hace que los *Q-TOF* sean muy útiles, ya que podemos aplicar rampas de energía de colisión a los iones en un tiempo muy corto (unas décimas de segundo). De este modo, si trabajamos adquiriendo dos funciones secuenciales, una de baja energía (unos 4 eV de energía de colisión) y otra de alta energía (típicamente, una rampa de energía de colisión entre 15 y 40 eV), en la primera adquisición (función) los iones no se fragmentan y podemos observar el ion pseudomolecular ($[M+H]^+$ ó $[M-H]^-$), mientras que en la segunda función se observan los fragmentos generados a partir de todos los iones que logran cruzar el filtro cuadrupolar. Debido a que los picos cromatográficos duran unos pocos segundos (en UHPLC alrededor de 6s) y las adquisiciones simultáneas se realizan varias veces por segundo (unos 0.3s por adquisición), en un pico de 6 segundos podremos tener 20 puntos de adquisición, 10 que corresponderán a la función de baja energía (ion sin fragmentar) y 10 puntos de alta energía (fragmentos), de manera que tenemos información de la molécula intacta e información estructural en un mismo análisis con suficientes puntos para definir el pico cromatográfico. Este modo de trabajo se denomina MS^E en los instrumentos de la casa comercial Waters, pero podemos encontrar otras nomenclaturas para otros fabricantes como Agilent (All Ions MS/MS) o Thermo (All Ions Fragmentation, AIF).

Finalmente, los iones (intactos o fragmentos) entran al analizador de tiempo de vuelo donde se miden sus relaciones *m/z* de manera exacta y con elevada resolución. En este tipo de análisis se obtiene una enorme cantidad de datos, haciendo necesaria la bioinformática a la hora de extraer la información requerida para los análisis metabolómicos. Además también proporciona información del patrón isotópico, extremadamente útil a la hora de establecer la composición elemental correcta de los iones observados.

El segundo modo de trabajo, que es extremadamente útil a la hora de la elucidación estructural, es el modo de barrido de iones producto (MS/MS). En este modo de trabajo, el cuadrupolo aísla un ion precursor en concreto, con una ventana de resolución cercana a 1 Da, la cual nos permite obtener información de un ion sin su patrón isotópico). De manera que a la celda de colisión entra una pequeña ventana de iones preseleccionados por el usuario. En la celda de colisión se aplica, en este caso, una energía de colisión concreta, que nos aporte información estructural de la molécula a energías de colisión altas y bajas. De este modo, cada pocos segundos una única relación *m/z* entra en la celda de colisión, se fragmenta, y sus iones producto son analizados en el tiempo de vuelo con información de masa exacta, obteniendo así desde información sobre las pérdidas lábiles (como pueden ser pérdidas de agua, de amoniaco,…) a energías de colisión bajas, hasta información de la estructura más "dura" de la molécula, de su esqueleto, a energías de colisión más altas (pérdida de ácidos grasos, iones tropilio,…).

De este modo, con un único instrumento somos capaces de realizar un barrido general para aplicar el flujo de trabajo metabolómico. En este punto, debido al modo de adquisición de los datos con alta y baja energía de fragmentación (MS^E), podemos hacer uso de la función de baja y alta energía para identificar la molécula si es posible. Si la información observada no fuera concluyente, bien por falta de fragmentación, coelución del pico cromatográfico u otro motivo, existe la posibilidad de realizar análisis de espectrometría de masas en tándem para elucidar su identidad aislando el ión precursor y observando un espectro limpio de interferencias. En este tipo de análisis, sin embargo, se tiene menor sensibilidad y, al ser análisis a posteriori, puede ser que la muestra haya podido sufrir transformaciones o degradaciones durante el tiempo de tratamiento de datos, a pesar de que por norma general, la información obtenida es mucho más concluyente.

Separación por movilidad iónica

En los últimos años ha aparecido una nueva dimensión de separación que se ha introducido en los instrumentos de espectrometría de masas, llamada Separación por Movilidad Iónica (IMS). Esta técnica basa su funcionamiento en la aplicación de una fuerza por medio de un campo eléctrico a un ion para así medir el tiempo que este tarda en cruzar una celda de movilidad (Gabelica & Marklund, 2018). Hay diferentes tipos de instrumentos que realizan esta función, como son los llamados "*drift tubes*" que fueron los primeros en aparecer o los "*travelling waves*" que aparecieron posteriormente. Todos ellos, tras su calibración, permiten relacionar el tiempo de deriva observado con el valor de sección de colisión transversal que presentan los iones al cruzar la celda de movilidad (Collisional Cross Section, CCS).

A pesar de la elevada cantidad de información que se tiene con la separación cromatográfica de alta resolución y la información de masa exacta, el conocimiento del CCS para un ión proporciona información adicional que puede resultar muy valiosa a la hora de obtener una identificación correcta (De Vijlder et al., 2017), o incluso unido a herramientas de predicción de CCS (Bijlsma et al., 2017) puede proporcionar información clave a la hora de identificar un compuesto candidato para el que no se dispone de patrón de referencia comercial.

I.3.3. Tratamiento de datos

Dentro del flujo de trabajo metabolómico, tras la preparación de muestra y la elección de la/s técnica/s analítica/s, sigue el tratamiento de datos. Esta parte del proceso requiere principalmente de la bioinformática y la quimiometría para su desarrollo. La quimiometría se define como el conjunto de herramientas matemáticas y estadísticas para su aplicación en química (Lavine & Workman, 2008) mientras que la bioinformática, además de esto, también abarca el almacenamiento, obtención, análisis e interpretación de la información generada desde un punto de vista biológico (Bains, 1996). A diferencia del procesamiento de datos en los estudios dirigidos, en los estudios no dirigidos, cada una de las etapas del flujo de trabajo debe ser optimizada, incluyendo el tratamiento de datos.

Para ello, el primer paso es convertir el formato privativo en el que los datos crudos son adquiridos a un formato genérico para este tipo de información, como puede ser *netCDF* o *mzXML*. Existen varios softwares para convertir los datos de los instrumentos a formato genérico, algunos incluidos en el propio software del fabricante (*DataBridge* en *MassLynx,* por ejemplo), o softwares disponibles en internet (*Proteowizard,* http://proteowizard.sourceforge.net/). Sin embargo, para determinados softwares, como puede ser UNIFI (Waters), que además almacena los datos no de manera individual para cada muestra sino en archivos que incluyen todo el set de muestras, no existe un conversor de datos capaz de adaptar la información de CCS, y debemos recurrir a software suministrado por el vendedor (Progenesis QI, Non-Linear Dynamics), puesto que trabaja directamente con los datos en formato propietario.

Una vez transformado el formato, el siguiente paso es la detección de especies iónicas (llamados "*features*"). Este paso resulta determinante a la hora de obtener un set de datos de buena calidad, ya que los parámetros a optimizar pueden llevar a eliminar gran parte de la información o a obtener un elevado número de falsos positivos. En este sentido, lo ideal es trabajar con un software que permita modificar parámetros de *peak picking* a la vez que permita visualizar los datos durante en el proceso, para así evitar este tipo de errores.

De entre todas las herramientas bioinformáticas disponibles gratuitamente en internet, la que se ha utilizado en esta Tesis Doctoral ha sido XCMS (Smith, Want, O'Maille, Abagyan, & Siuzdak, 2006), un paquete escrito en lenguaje de programación R (http://www.rproject.org/) enfocado al tratamiento de datos de cromatografía-espectrometría de masas que permite realizar varios procesos, descritos a continuación, para obtener finalmente una tabla de datos con el área integrada de los *features*. Los procesos generales del tratamiento de datos en metabolómica son los siguientes:

Peak picking

En este proceso el software extrae de cada muestra los *features* . El algoritmo utilizado en el desarrollo de esta tesis, entre los disponibles en XCMS, ha sido *CentWave*. Este algoritmo funciona buscando, para cada relación *m/z* con un error determinado (15 ppm normalmente, ya que las secuencias largas pueden modificar la exactitud de masa del *Q-TOF*), los picos cromatográficos que aparezcan a esta relación *m/z*. Las variables de selección se pueden modificar por el usuario. En el caso que nos ocupa se establecieron como estándares que definen un pico cromatográfico un tiempo entre 4 y 20 segundos, al menos 3 puntos por encima de 1000 cuentas y con una relación señal/ruido superior a 10. Cada uno de estos picos generan una *feature*, para la cual se integra el área, y que se etiqueta como M*xxx*T*yyy*, donde *xxx* es la masa nominal e *yyy* el tiempo de retención en segundos para cada *feature*.

Alineación de tiempos de retención

Este paso es crucial, ya que la variación de un segundo en el tiempo de retención de un ion de una muestra a otra hace que la etiqueta que se le ha asignado no sea la misma (M254T225 en una muestra, M254T226 en la siguiente…), a pesar de que el ion sí que lo sea (*m/z* 254.2456, por ejemplo). Por este motivo, se realiza una alineación de tiempos de retención entre muestras, para unir los *features* que, aun correspondiendo al mismo ion, se han etiquetado como diferentes. La ventana de corrección depende mucho de cada experimento, por lo que será necesario una visión previa a nuestras muestras para determinar la variación instrumental de tiempos de retención que hemos tenido entre las primeras muestras y las ultimas en un set de muestras, aunque este no debería exceder en ningún caso de los 15 segundos. En esta Tesis Doctoral se ha utilizado la función *group* de *XCMS*, con una corrección de tiempos reduciéndolos desde los 15 segundos (bw=15) hasta menos de un segundo.

Rellenado de picos (*FillPeaks*)

Este paso resulta muy útil para obtener una matriz de resultados mucho más ajustada a la realidad, debido a que los picos en los que el área es 0, probablemente no hayan sido integrados a causa de las restricciones del primer paso (Peak Picking). Con este rellenado de picos, ese 0 se

sustituye por la respuesta real del compuesto en la muestra, haciendo que diferencias que podían parecer realmente importantes (un compuesto que en un grupo se encuentre ligeramente por encima del punto de corte y en otro ligeramente por debajo) en realidad no lo sean.

Normalización

Este paso es recomendable debido a las posibles derivas instrumentales que hayan podido surgir por el uso continuado del equipo (posibles caídas de señal a lo largo de la secuencia, suciedad acumulada en el cono de extracción, etc). Esta caída se puede observar claramente en el grupo de muestras llamado QC (**Figura I.12**), las cuales se han definido en la sección de *tratamiento de muestra*. Esta muestra se genera con el objetivo de controlar el proceso de normalización, ya que se inyecta a lo largo de la secuencia alternativamente a un número de inyecciones (en nuestro caso 10) y, al tratarse de la misma muestra, debería tener la misma señal a lo largo de todo el set. En caso de que no sea así, será necesaria la normalización. Como se puede observar en la **Figura I.12**, las muestras QC presentaban una caída de señal en los datos crudos (izquierda), mientras que en los datos normalizados (derecha) esa caída se ha corregido.

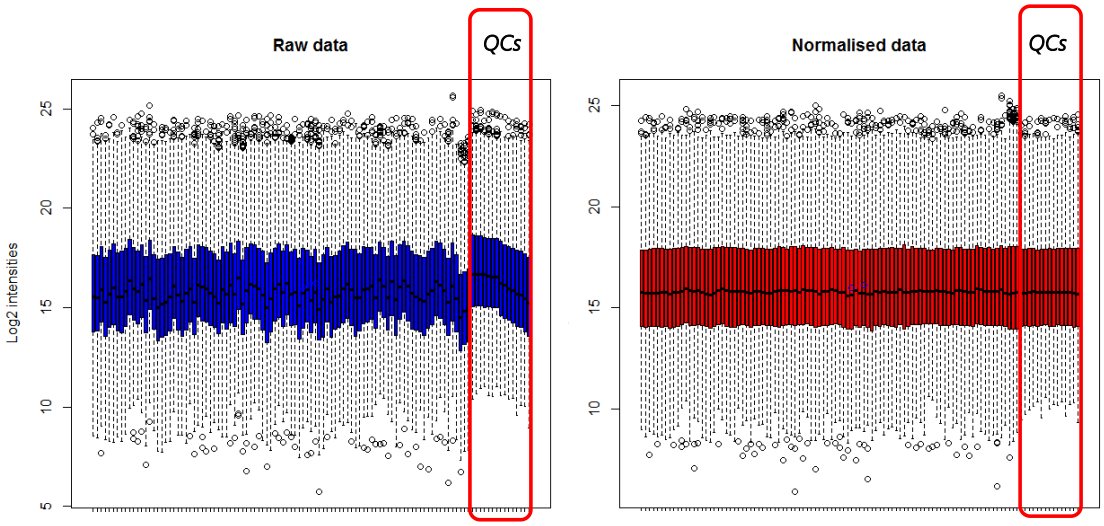


Figura I.12: Comprobación de la correcta normalización del set de muestras debido al agrupamiento de las muestras QC en la misma intensidad mediana.

Las posibles normalizaciones que se pueden utilizar se pueden consultar en la literatura (van den berg et al, 2006). En nuestro caso, se ha aplicado el centrado a la mediana para obtener

una normalización del set completo de muestras. Posteriormente, para eliminar la heteroscedasticidad, se aplica el logaritmo en base 2 a las respuestas. Esto tiene un efecto de escalado ligero, aunque no suficiente para ajustar las diferencias entre los analitos que, sin embargo, es reversible. Finalmente, para reducir la desviación estándar a 1 en todos los analitos y así destacar los compuestos de interés independientemente de su señal y desviación se aplica un escalamiento de los datos que, en nuestro caso, fue *Pareto scaling*. Este escalamiento es capaz de reducir la importancia de valores muy grandes manteniendo la estructura de los datos casi intacta y siendo fiel por tanto a los datos originales.

Sin embargo, en sets de muestras pequeños, con muy poca deriva en el instrumento, en los que no se observa caída alguna en los QC, o en sets de muestra heterogéneos, es aconsejable evitar el centrado, ya que cualquier modificación de los datos puede esconder o generar diferencias que, al intentar validar, se podrían volver inexistentes. Por otro lado, el escalado, al no afectar tan internamente a los datos, puede ser aplicado.

Como ya se ha expuesto anteriormente, en datos obtenidos con cuatro dimensiones (RT, *m/z*, CCS y área), el único programa que hasta la fecha puede ser utilizado es *Progenesis QI*. El software resulta muy fácil de utilizar y dinámico. Permite revisar los pasos de deconvolución (que no así escoger métodos o parámetros en ellos), y es capaz de trabajar con datos de diferentes instrumentos, permitiendo realizar una búsqueda de identidades en Chemspider e incluso facilita trasladar los datos a un software de análisis estadístico (EZ-Info, Umetrics). Sin embargo, como punto débil se podría poner que el software permite modificar pocos parámetros en comparación con XCMS, además de no ser gratuito.

Transcripción de la matriz de datos

Finalmente, los datos obtenidos se introducen en una matriz con información de masa exacta, tiempo de retención, CCS (si la hay) y área de los compuestos en cada muestra, entre otros, que ya está lista para ser analizada estadísticamente.

I.3.4. Tratamiento estadístico

En metabolómica se han utilizado desde su aparición diferentes métodos estadísticos que, aún a día de hoy, se actualizan o surgen nuevos. Sin embargo, todos ellos se pueden dividir en dos grandes bloques, los métodos estadísticos "univariantes", que relacionan únicamente la información de una variable (*feature*), y los "multivariantes" que lo hacen de manera conjunta con la información de más de una variable para todas las muestras seleccionadas.

Análisis estadísticos univariantes

El análisis de la varianza (ANOVA) es un estadístico univariante que se utiliza para conocer si los grupos de estudio (tres o más grupos) son o no significativamente diferentes, con una probabilidad de errar generalmente menor al 5%. Cuando se observan medias significativamente distintas, se aplica la prueba de t *de student* para conocer en detalle que grupos tienen medias distintas y cuáles no.

Sin embargo, debido a la enorme cantidad de datos con los que trabajamos en metabolómica, es necesaria la corrección de los valores proporcionados por ANOVA para así evitar los falsos positivos, ya que por puro azar un porcentaje de variables aparecen como significativamente distintas sin serlo realmente. Para ello se utiliza la estimación de "*False Discovery rate* " de Benjamini-Hockberg. Esta herramienta puede ser aplicada de diversos modos, existiendo funciones en MATLAB (*FDR*) o incluso páginas web gratuitas que permiten realizar este cálculo utilizando los *p-value* (https://www.sdmproject.com/utilities/?show=FDR). De este modo se realiza un análisis más restrictivo, reduciendo el número de falsos positivos (marcadores que aparecen significativamente diferentes sin serlo) a pesar de que ello provoque generar falsos negativos (perder marcadores significativos debido a la corrección restrictiva).

Análisis estadísticos multivariantes

En este grupo se encuentran los métodos de clasificación "no supervisados" y los "supervisados". Entre ellos la única diferencia radica en el uso que hacen de la información de pertenencia de las muestras a los grupos preestablecidos.

De este modo, los métodos "no supervisados", al no tener en cuenta la pertenencia a ningún grupo, proporcionan información sobre la relación que las muestras tienen entre sí de manera general (parecidos entre grupos, etc…).

PCA

Entre ellos se encuentra el Análisis de Componentes Principales (PCA), que se ha utilizado a lo largo de la tesis con el objetivo fundamental de confirmar la correcta normalización de las muestras, así como para observar patrones de comportamiento y/o detectar muestras atípicas, llamadas "*outliers*".

Este método se basa en la reducción de una enorme cantidad de variables, como es el caso de los experimentos de cromatografía-espectrometría de masas, a un nuevo sistema de coordenadas generado por los llamados componentes principales (PCs). Estas componentes principales no son sino combinaciones lineales de las variables introducidas al modelo, de modo que se generan combinaciones para intentar explicar la máxima varianza posible entre las muestras. Finalmente, el modelo reduce la información que existe de las muestras (dada generalmente por más de 1000 *features*), a unas pocas variables o PCs (generalmente menos de 10), explicando un tanto por cien de la varianza real de las muestras. Cuanto mayor porcentaje se alcanza a explicar, mejor representan estas componentes a nuestro set de muestras.

A partir de este análisis es necesario, por un lado, comprobar como las muestras QC se agrupan en un punto centrado en el Plot (ver **Figura I.13**) y por otro lado, es importante identificar y eliminar los "outliers" para evitar distorsiones de los posteriores análisis estadísticos. Las muestras QC, como ya se ha expuesto anteriormente, al ser básicamente la misma muestra inyectada de manera consecutiva, debe tener las mismas coordenadas dentro de nuestro nuevo sistema de componentes principales, a la vez que encontrarse cercanos al origen de coordenadas (0,0), lo que nos indica que los pasos de tratamiento de datos (básicamente la normalización) han funcionado correctamente.

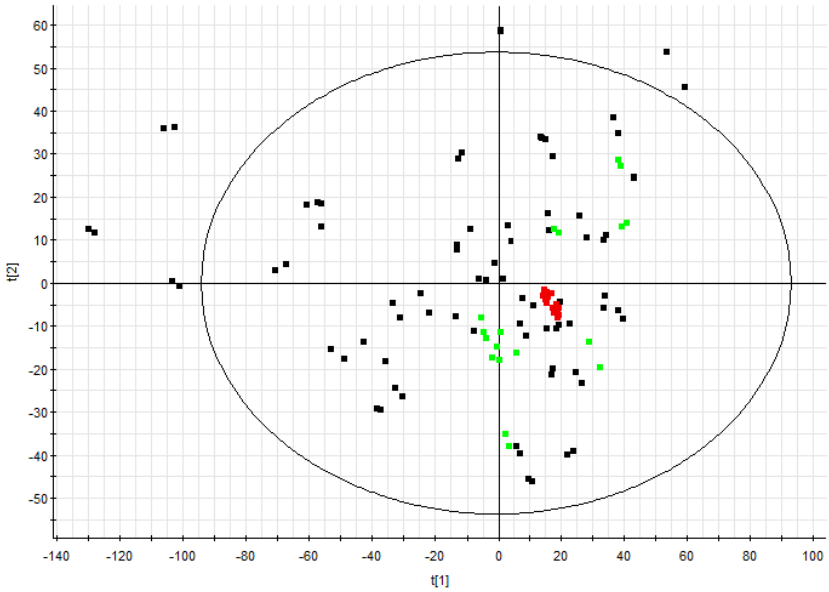


Figura I.13: PCA aplicado a un grupo de muestras heterogéneo. En rojo, las muestras QC inyectadas al inicio de la secuencia y a lo largo de ésta.

Es también importante identificar los "*outliers*", sabiéndolos diferenciar de muestras que, simplemente por su composición, son diferentes del resto. Para ilustrar esto se puede observar la **Figura I.13** de nuevo. A pesar de que hay 6 puntos que, a priori, pueden parecer outliers (en la zona izquierda del Score Plot), realmente son muestras con una composición muy diferente, debido a que en el citado experimento el grupo definido con puntos negros presenta una heterogeneidad muy elevada. Por lo tanto, estos puntos no deben ser descartados del experimento. Esta heterogeneidad en las muestras, a pesar de no demasiado aconsejable en metabolómica, puede aportar marcadores muy robustos si los hay, ya que definen a un grupo de muestras que *a priori* podrían resultar difíciles de unificar.

Por otro lado están los métodos "supervisados", como pueden ser Partial Least Squares-Discriminant Analysis (PLS-DA), Orthogonal Partial Least Squares – Discriminant Analysis (OPLS-DA). Estos métodos son útiles para extraer información entre dos (OPLS-DA) o más grupos (PLS-DA). También existen otros modelos estadísticos, más utilizados para autentificación, como puede ser el Data Driven Soft Independent Modelling of Class Analogy (DD-SIMCA), entre otros. Los

métodos citados han sido los escogidos en los trabajos de esta tesis doctoral, por lo que van a ser explicados en más detalle.

En este punto cabe distinguir cual es el objetivo principal de los mismos, es decir, si se utilizan para observar y remarcar diferencias entre grupos o si su utilidad es para crear un modelo con las variables seleccionadas y aplicarlo en muestras reales. En este sentido, PLS-DA y OPLS-DA cumplen con el primer objetivo.

PLS-DA

Este método estadístico, muy similar en su modo de trabajo al PCA, crea unas nuevas coordenadas para el set de muestras, llamadas variables latentes (LVs). Estas variables, que también son combinaciones lineales de las *features*, se diferencian de las PCs del PCA en que el PLS-DA utiliza un vector con información sobre la pertenencia de las muestras a determinados grupos (vector de clase). De este modo, las LVs se generan de manera que sean capaces de explicar la máxima varianza posible entre los grupos seleccionados mientras que las PCs explican la máxima varianza posible en el set de muestras (Ballabio & Consonni, 2013). Este nuevo hiperespacio generado debe ser capaz, de una manera visual, de agrupar y separar las muestras, como puede observarse en la **Figura I.14** parte izquierda. De este modo, el PLS-DA resulta extremadamente útil para seleccionar cuales son las mejores variables que explican la diferenciación de las muestras en los grupos preseleccionados. Para ello se utiliza el llamado Parámetro de Importancia en la Varianza (VIP). Este se trata de un número que define el peso de cada *feature* en cada una de las LVs, de modo que, si las LVs explican el 95% de la varianza, los *features* que más influencian en esta serán los que tengan un valor de VIP mayor. Una vez reducidos y seleccionados, se comprueba como realmente el número total de *features* analizadas en las muestras (que generalmente supera las 1000) se pueden reducir a menos de 50, en la mayoría de los casos, y todavía clasificando correctamente las muestras en los grupos.

Una vez seleccionadas las variables que mejor explican los grupos, se deben validar las LVs. Este procedimiento se debe realizar en varios pasos, como se ha expuesto en la literatura (Riedl, Esslinger, & Fauhl-Hassek, 2015), con una validación interna por medio de *cross-validation* y una validación externa con muestras del mismo año y de otros años para establecer la validez de

estos en modelos de autentificación de alimentos (**Figura I.14** parte derecha), o con muestras reservadas del mismo experimento y experimentos repetidos en el caso de experimentos biológicos, aunque una validación más exhaustiva se puede obtener con la interpretación biológica de los resultados.

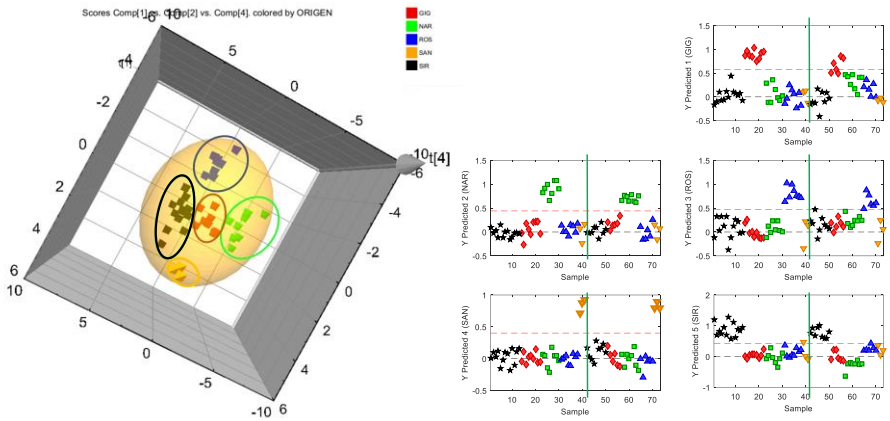


Figura I.14: A la izquierda score plot de un PLS-DA donde se observa la posición de cada una de las muestras en el nuevo hiperespacio, y a la derecha predicción de las muestras utilizadas para crear el modelo (izquierda de la línea verde) y para validarlo (derecha de la línea verde).

OPLS-DA

En el caso del OPLS-DA, también similar a PLS-DA, se trabaja de forma que las componentes se generan para separar dos grupos preseleccionados. La primera componente (predictiva) es la encargada de separar los grupos, mientras que la segunda componente (ortogonal) se encarga de explicar la varianza dentro de estos (Pinto, Trygg, & Gottfries, 2012). En este caso, OPLS-DA se utiliza para conocer los *features* que, de forma individual, son mejores para separar los dos grupos. Para ello se representa el llamado S-Plot (**Figura I.15**). En este, cada uno de los grupos se sitúa en el eje Y, siendo el grupo 1 la parte superior del gráfico y el grupo 2 la parte inferior. De este modo, el P[corr] representado en el eje Y, indica la correlación del *feature* con el grupo que mejor representa, de modo que los valores cercanos a 1 significan que la correlación del marcador con el grupo 1 es muy alta, mientras que -1 indica que el marcador está más relacionado con el grupo 2. En el eje X, por su parte, se representa la covarianza del

compuesto, de manera que los compuestos más a la derecha serán compuestos con una magnitud mayor. Así pues, como se observa en la **Figura I.15**, los mejores marcadores para separar entre ambos grupos serán los de la zona verde.

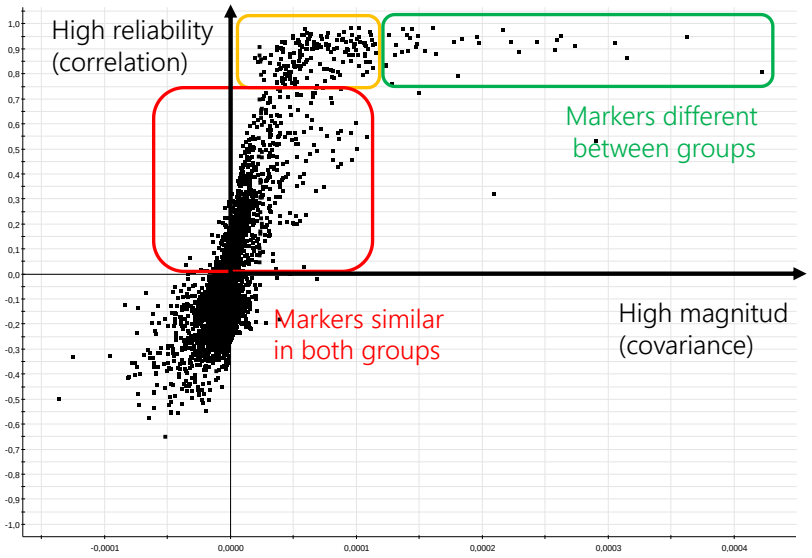


Figura I.15: Gráfico S-plot donde se observa la zona en la que se deben seleccionar los marcadores.

Este tipo de modelo estadístico resulta, por lo tanto, extremadamente útil para separar dos grupos, y presenta la característica frente a modelos univariantes de que no solo tiene en cuenta la diferencia entre ambos grupos sino también la magnitud de los *features*.

Stepwise Variable Selection Model (SVSM)

Además de VIP en PLS-DA o el S-Plot en OPLS-DA, se han utilizado en la tesis doctoral otros métodos de selección de variables como el SVSM. En este caso, se trata del artículo científico 1, donde se utilizó para comprobar como las variables seleccionadas con el S-Plot de OPLS-DA resultaban satisfactorias en la creación del modelo de clasificación. Este modelo, en su modo de trabajo de eliminación *Backward elimination*, elimina de uno en uno los marcadores que superen el p-valor establecido como crítico y recalcula el modelo, repitiendo el anterior paso hasta que ninguno de ellos supera el valor crítico, lo que implica que el modelo pierde consistencia. Sin

embargo, además de utilizarse como método de reducción de variables, se puede utilizar para confirmar que ninguno de los compuestos seleccionados por otras vías supera este valor y confirmando, por lo tanto, la necesidad de que todos ellos sean utilizados en el modelo.

A pesar de que resultó útil como validación del método de selección de variables basado en OPLS-DA, este método se decidió sustituir por la selección de variables VIP del PLS-DA ya que resulta mucho menos costoso en cuanto a tiempo, por lo que se utilizó únicamente en el artículo científico 1.

DD-SIMCA

En este caso, DD-SIMCA se trata de un método estadístico de autentificación de una clase. Esto significa que su funcionamiento se basa en determinar si una muestra pertenece (o no) a dicho grupo. Si es así se etiqueta como tal y en caso contrario, se etiqueta como desconocida. La diferencia principal respecto a PLS-DA es que DD-SIMCA funciona bien con muestras desconocidas (llamadas "alien") no clasificándolas erróneamente (falsos positivos), mientras que PLS-DA solo funciona correctamente con muestras incluidas en el modelo, incurriendo en este tipo de errores. Este hecho hace que DD-SIMCA sea un modelo estadístico correcto para la creación de métodos de autentificación, mientras que PLS-DA se debe utilizar para selección de variables y no para autentificación, como ha sido publicado recientemente en la literatura (Rodionova, Titova, & Pomerantsev, 2016; Zontov, Rodionova, Kucheryavskiy, & Pomerantsev, 2017) y será demostrado y comentado ampliamente en el capítulo III de esta tesis.

DD-SIMCA trabaja basado en un modelo PCA, con el cual descompone la matriz de datos en PCs, para en un segundo paso calcular la distancia de la muestra al grupo (SD, *score distance*), llamada h_i, y la distancia ortogonal (OD, *orthogonal distance*), llamada v_i. Posteriormente, se define una variable de aceptación, con la que las muestras utilizadas para generar el modelo queden dentro y, a su vez, las muestras utilizadas para validarlo también sean correctamente clasificadas.

De este modo, tenemos gráficos como el mostrado en la **Figura I.16**, donde se puede observar como las muestras utilizadas para generar el modelo (figura de la izquierda) entran dentro de la zona de aceptación, al igual que las muestras utilizadas para validarlo (figura de la

derecha), mientras que el resto de muestras del ejemplo quedan fuera de este grupo y no serían etiquetadas, en este caso, como GIG.

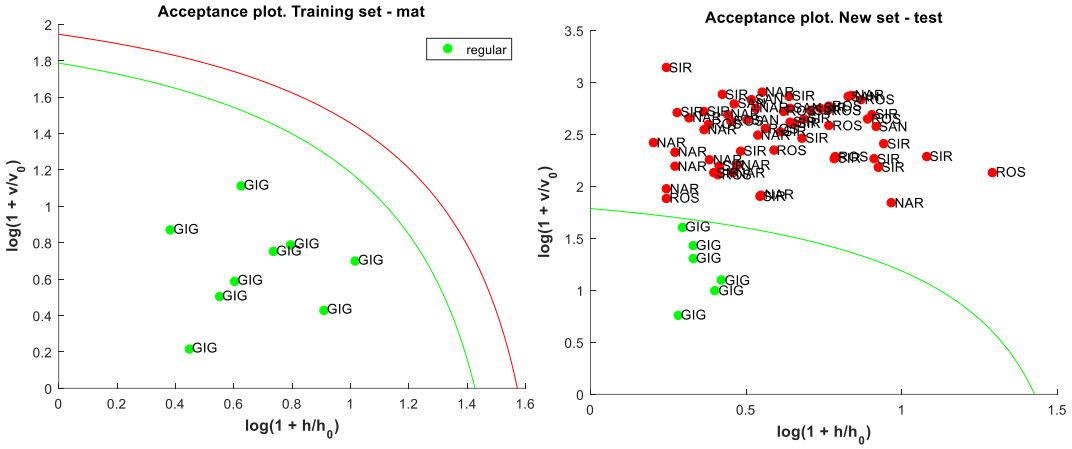


Figura I.16: Gráficos de aceptación en DD-SIMCA. A la izquierda, con las muestras utilizadas para crear el modelo. A la derecha, con los resultados de las muestras utilizadas para validarlo.

I.3.5. Elucidación estructural

Una vez seleccionadas las *features* que nos permiten obtener una buena clasificación por medio de los métodos estadísticos, el siguiente paso, dadas las posibilidades que arroja la espectrometría de masas de alta resolución en este campo, es la elucidación estructural de los compuestos seleccionados. Este proceso resulta ser el que más tiempo conlleva en metabolómica no dirigida, ya que los compuestos se desconocen y, además, se desconoce cuál puede ser su naturaleza (*unknown unknowns*) (Wishart, 2009). En el caso de las matrices biológicas, se puede reducir bastante el rango de compuestos a metabolitos biológicos, a pesar de que siempre se puede observar algún compuesto inesperado, mientras que en el caso de matrices de alimentos o ambientales el rango de compuestos en el que se trabaja puede parecer inabarcable. Sin embargo, en esta tesis se ha tratado de definir un flujo de trabajo para intentar conseguir resultados satisfactorios, que se enumera a continuación:

Definir una composición molecular

El primer paso, tras la selección del compuesto por medio de los métodos estadísticos, siempre es establecer la fórmula molecular del compuesto. En este punto nos encontramos con la indiscutible necesidad de trabajar en metabolómica basada en el acoplamiento LC-MS con espectrómetros de masas de alta resolución, ya que debemos reducir la enorme cantidad de posibles composiciones elementales a una en concreto, la que mejor coincida tanto con la masa exacta como con el patrón isotópico de nuestro compuesto. Sin embargo resulta importante conocer primero si nuestro *feature* es el ion pseudomolecular ($[M+H]^+$/ $[M-H]^-$) o un aducto de éste con otros contraiones ($[M+Na]^+$,$[M+K]^+$, $[M+Cl]^-$, $[M+HCOO]^-$). En este sentido es importante observar el espectro de masas de baja energía de colisión proporcionado por el modo MS^E, donde buscaremos estas diferencias de *m/z* en la función de baja energía de colisión (+21.982 para el aducto con Na^+ respecto al ion protonado, 17.0265 para el aducto con NH_4^+,…) o incluso, en casos en que el ion protonado no se observe, diferencias entre aductos (+4.9555 entre aductos de NH_4^+ y Na^+,…). En este sentido, debemos esclarecer cual es la masa del ion pseudomolecular para no incurrir en errores a la hora de elucidar nuestro compuesto, asignando elementos que simplemente aparecen como contraiones.

Posteriormente, teniendo en cuenta las abundancias isotópicas de los elementos, podemos reducir estos patrones a posibles composiciones moleculares (por ejemplo, la presencia de un cloro en la estructura generará una señal a 1.997 Da del ion pseudomolecular seleccionado del 33% debido a las abundancias relativas del cloro (^{35}Cl:^{37}Cl = 75:25) mientras que la presencia de un azufre en la estructura solamente genera un pico de intensidad 4.4 % a 1.996 Da (^{32}S:^{34}S = 95.02:4.22)) y la presencia de carbono, hidrógeno, oxígeno o nitrógeno apenas se ve reflejada a no ser que la cantidad de estos sea muy elevada.

De este modo, llegamos al primer punto del proceso, establecer la composición elemental de la fórmula molecular. Sin embargo, esta composición puede corresponder a muchas estructuras químicas, por lo que el proceso de elucidación no acaba aquí.

Estudio de los patrones de fragmentación

En este punto es importante observar como el compuesto se fragmenta, para intentar relacionar los iones producto con una estructura plausible, que pueda explicarse a través de la molécula padre. Para ello puede resultar útil la función de alta energía del MS^E, aunque por norma general, esta solo aporta información si el ion marcador es mayoritario en su zona de tiempo de retención. En caso contrario, en extremadamente beneficioso para el proceso de elucidación utilizar el modo de masas en tándem (MS/MS) que nos permiten los instrumentos híbridos, como el *Q-TOF*, ya que podemos aislar el ion de interés a través del primer cuadrupolo, fragmentarlo a distintas energías de colisión y posteriormente analizar sus iones producto para intentar esclarecer cual es la estructura del compuesto sin incurrir en errores al asignar fragmentos a las relaciones *m/z* observadas.

Una vez obtenida toda esta información, se debe proseguir estudiando estos patrones de fragmentación. En este punto, se pueden observar varias pérdidas neutras que podrán guiarnos a la hora de seleccionar una estructura u otra, como puede ser pérdidas neutras de agua (H_2O, -18.0105 Da), metilo (CH_3, -15.0234 Da), amoniaco (NH_3, -17.0265 Da), de un grupo ciano (HCN, -27.0109 Da), un ácido fórmico en modo de ionización positivo (HCOOH, -46.0054 Da), dióxido de carbono (CO_2, -43.9898 Da) en negativo o incluso de cadenas más largas como puede ser una hexosa ($C_6H_{12}O_6$, -180.0630 Da). Toda esta información nos aporta subestructuras que deberá contener nuestra molécula. Con esta información podemos llegar a una elucidación tentativa del compuesto. Sin embargo, normalmente no tendremos suficiente información para poder asegurar de qué molécula en concreto se trata, por lo que tenemos que recurrir a la búsqueda en bases de datos.

Búsqueda en bases de datos

Hay diferentes tipos de bases de datos a las cuales podemos recurrir para obtener una identificación del compuesto, desde bases de datos generales (*ChemSpider, PubChem,…*) donde cualquier compuesto, descrito o no, podría encontrarse, hasta bases de datos más específicas. Estas pueden ser bases de datos de metabolitos (*mzCloud*, *Metlin*, *HMDB*,..), lípidos y compuestos relacionados (*LipidMaps*) o incluso de fármacos, pesticidas… (*MassBank*). Dependiendo del tipo de muestra con el que trabajemos, deberemos seleccionar una o varias de ellas para la búsqueda de candidatos para el marcador.

Una vez seleccionamos la base de datos más adecuada, podremos buscar espectros de masas en tándem para compararlos con nuestros marcadores, si se encuentran disponibles, o simplemente obtener posibles estructuras que expliquen los espectros obtenidos. En caso de que no se obtenga ningún resultado en bases de datos específicas, se puede recurrir al uso de "predictores" de espectros de masas *in-silico*, entre los que se encuentra *MetFrag*. En este tipo de programas, disponibles gratuitamente en internet, podemos realizar una búsqueda del compuesto por composición elemental o incluso por relación *m/z* en caso de dudar entre dos o más de ellas, añadiendo información acerca de los fragmentos observados en el espectro. Con todo ello, el software obtiene información de bases de datos generales (*Chemspider, PubChem,…*) sobre las moléculas que se adapten a esta composición elemental o *m/z*, e intenta explicar los iones producto observados a partir de la estructuras químicas de los candidatos, obteniendo finalmente un listado de los compuestos que mejor se adaptan a la búsqueda.

Si llegados a este punto no hemos obtenido ningún resultado satisfactorio, sin conocer siquiera su composición elemental, daremos la molécula como no identificada, o de *nivel 1* (Schymanski et al., 2014). Si llegamos a una formula molécular posible, pero no a una identificación inequívoca, diremos que es *nivel 2*. Si por el contrario, llegamos a obtener una elucidación tentativa, deberemos seguir con el proceso de elucidación.

Identificación con un patrón comercial

En caso de tener un candidato, el siguiente paso es obtener un patrón comercial del producto para comparar su espectro de fragmentación, su tiempo de retención y, en su caso, el valor de CCS. Si el patrón está disponible y todos los parámetros coinciden, tenemos una identificación de *nivel 5* (estructura confirmada), en caso de que no esté comercial, podemos utilizar herramientas de predicción del tiempo de retención (Bade et al., 2015) y de CCS (Bijlsma et al., 2017). Si estos parámetros se acercan a los parámetros observados, tendremos una identificación de *nivel 4*, ya que la confianza en nuestra elucidación aumentará en gran medida. Si el error en estos parámetros es alto, tendremos una identificación de *nivel 3*.

I.3.6. Interpretación biológica

Una vez obtenida la identificación tentativa de la/s molécula/s de interés es posible continuar con su interpretación biológica. En el caso de crear métodos para autentificación, este paso proporciona una información útil acerca de la identidad de los compuestos seleccionados, para así evitar que nuestros métodos utilicen xenobióticos u otros compuestos que aparecen de forma exógena en las muestras. Sin embargo, en los experimentos de metabolómica de organismos biológicos es casi más importante este proceso que la propia identificación, hecho ilustrado en los artículos científicos del capítulo II.

Existen para elllo bases de datos en internet con información de rutas metabólicas disponibles gratuitamente, como *KEGG* (Kyoto Encyclopedia of Genes and Genomes), que permiten relacionar compuestos entre sí, obteniendo no solo información de los compuestos exaltados por la metabolómica no dirigida, sino aportando información biológica sobre los marcadores observados. Sin embargo, dadas las dificultades que presenta la identificación de compuestos desconocidos y la enorme cantidad de ellos que se pueden obtener en experimentos donde el cambio entre los grupos sea muy grande (por ejemplo el artículo científico 1), se debe realizar una criba para seleccionar y elucidar los iones que resultan más representativos de estas modificaciones metabólicas. Es decir, seleccionar los marcadores que, de una forma más robusta,

nos ayuden a explicar y controlar posteriormente el cambio observado. Dada la capacidad que nos proporciona el modo de adquisición Full scan en los equipos *Q-TOF*, podemos volver a los datos, en el llamado análisis retrospectivo, a buscar compuestos relacionados (pertenecientes al mismo metabolic pathway alterado) con los que la estadística ha marcado y así observar su comportamiento. Sin embargo, sin tener un patrón de referencia para asegurar que el compuesto que queremos observar se encuentra en la muestra, resulta imprescindible el modo de adquisición MS^E, donde además del ion pseudomolecular, podemos realizar nwXICs (narrow window extracted ion chormatograms) de los fragmentos esperados para el compuesto y, de este modo, confirmar la identidad de estos compuestos relacionados de una manera más inequívoca.

Estos compuestos obtenidos en una segunda búsqueda, a la que llamamos "refining process", resultan extremadamente útiles para comprender la forma en que los "pathways" metabólicos se han visto afectados, proporcionando seguridad a nuestra interpretación biológica.

Cabe destacar, finalmente, que debido a la enorme dificultad que conllevan los procesos tanto de selección, de elucidación como de interpretación biológica, es extremadamente aconsejable trabajar con grupos de trabajo multidisciplinares (estadísticos, químicos, biólogos, médicos,…) para asegurar no solo que los resultados se han obtenido de una manera correcta, sino que la interpretación de estos también se ha tomado en consideración por especialistas. Además, la unión de las tecnologías ómicas (genómica, transcriptómica, proteómica y metabolómica) en una aproximación "holística" siempre proporciona una mejor comprensión de los resultados, así como una mayor seguridad en ellos (Fiehn, 2001) .

Referencias

Adam, A.-C., Lie, K. K., Moren, M., & Skjærven, K. H. (2017). High dietary arachidonic acid levels induce changes in complex lipids and immune-related eicosanoids and increase levels of oxidised metabolites in zebrafish (Danio rerio). *British Journal of Nutrition*, *117*(8), 1075–1085.

Altelaar, A. F. M., Munoz, J., & Heck, A. J. R. (2012). Next-generation proteomics: towards an integrative view of proteome dynamics. *Nature Reviews Genetics*, *14*(1), 35–48.

Andersen, M.-B. S., Reinbach, H. C., Rinnan, Å., Barri, T., Mithril, C., & Dragsted, L. O. (2013). Discovery of exposure markers in urine for Brassica-containing meals served with different protein sources by UPLC-qTOF-MS untargeted metabolomics. *Metabolomics*, *9*(5), 984–997.

Andersen, S. M. (2015). Metabolomic analysis of plasma and liver from surplus arginine fed Atlantic salmon. *Frontiers in Bioscience*, *7*(1), 718.

Ardrey, R. E. (2003). *Liquid Chromatography – Mass Spectrometry: An Introduction*. Chichester, UK: John Wiley & Sons, Ltd.

Bade, R., Bijlsma, L., Miller, T. H., Barron, L. P., Sancho, J. V., & Hernández, F. (2015). Suspect screening of large numbers of emerging contaminants in environmental waters using artificial neural networks for chromatographic retention time prediction and high resolution mass spectrometry data analysis. *Science of the Total Environment*, *538*, 934–941.

Bains, W. (1996). Company strategies for using bioinformatics. *Trends in Biotechnology*, *14*(8), 312–317.

Ballabio, D., & Consonni, V. (2013). Classification tools in chemistry. Part 1: linear models. PLS-DA. *Analytical Methods*, *5*(16), 3790.

Barreira, J. C. M., Casal, S., Ferreira, I. C. F. R., Peres, A. M., Pereira, J. A., & Oliveira, M. B. P. P. (2012). Supervised chemical pattern recognition in almond (Prunus dulcis) Portuguese PDO cultivars: PCA- and LDA-based triennial study. *Journal of Agricultural and Food Chemistry*, *60*(38), 9697–9704.

Baumgarner, B. L., & Cooper, B. R. (2012). Evaluation of a tandem gas chromatography/time-of-flight mass spectrometry metabolomics platform as a single method to investigate the effect of starvation on whole-animal metabolism in rainbow trout (Oncorhynchus mykiss). *The Journal of Experimental Biology*, *215*(Pt 10), 1627–32.

Beltrán Sanahuja, A., Ramos Santonja, M., Grané Teruel, N., Martín Carratalá, M. L., & Garrigós Selva, M. C. (2011). Classification of almond cultivars using oil volatile compound determination by HS-SPME-GC-MS. *JAOCS, Journal of the American Oil Chemists' Society*, *88*(3), 329–336.

Benedito-Palos, L., Calduch-Giner, J. A., Ballester-Lozano, G. F., & Pérez-Sánchez, J. (2013). Effect of

ration size on fillet fatty acid composition, phospholipid allostasis and mRNA expression patterns of lipid regulatory genes in gilthead sea bream (Sparus aurata). *The British Journal of Nutrition*, *109*(7), 1175–87.

Bijlsma, L., Bade, R., Celma, A., Mullin, L., Cleland, G., Stead, S., … Sancho, J. V. (2017). Prediction of Collision Cross-Section Values for Small Molecules: Application to Pesticide Residue Analysis. *Analytical Chemistry*, *89*(12), 6583–6589.

Bino, R. J., Hall, R. D., Fiehn, O., Kopka, J., Saito, K., Draper, J., … Sumner, L. W. (2004, September 1). Potential of metabolomics as a functional genomics tool. *Trends in Plant Science*. Elsevier Current Trends.

Bloszies, C. S., & Fiehn, O. (2018). Using untargeted metabolomics for detecting exposome compounds. *Current Opinion in Toxicology*, *8*, 87–92.

Bruce, S. J., Tavazzi, I., Parisod, V., Rezzi, S., Kochhar, S., & Guy, P. A. (2009). Investigation of human blood plasma sample preparation for performing metabolomics using ultrahigh performance liquid chromatography/mass spectrometry. *Analytical Chemistry*, *81*(9), 3285–3296.

Brunella Cavaliere, Antonio De Nino, Fourati Hayet, Aida Lazez, Barbara Macchione, Cossentini Moncef, … Tagarelli, A. (2007). A Metabolomic Approach to the Evaluation of the Origin of Extra Virgin Olive Oil: A Convenient Statistical Treatment of Mass Spectrometric Analytical Data.

Cajka, T., Danhelova, H., Vavrecka, A., Riddellova, K., Kocourek, V., Vacha, F., & Hajslova, J. (2013). Evaluation of direct analysis in real time ionization–mass spectrometry (DART–MS) in fish metabolomics aimed to assess the response to dietary supplementation. *Talanta*, *115*, 263–270.

Calduch-Giner, J. A., Sitjà-Bobadilla, A., & Pérez-Sánchez, J. (2016). Gene Expression Profiling Reveals Functional Specialization along the Intestinal Tract of a Carnivorous Teleostean Fish (Dicentrarchus labrax). *Frontiers in Physiology*, *7*, 359.

Capozzi, F., & Bordoni, A. (2013). Foodomics: a new comprehensive approach to food and nutrition. *Genes & Nutrition*, *8*(1), 1–4.

Castro-Puyana, M., Pérez-Míguez, R., Montero, L., & Herrero, M. (2017). Application of mass spectrometry-based metabolomics approaches for food safety, quality and traceability. *TrAC Trends in Analytical Chemistry*, *93*, 102–118.

Cevallos-cevallos, J. M., Etxeberria, E., Danyluk, M. D., & Rodrick, G. E. (2009). Metabolomic analysis in food science : a review. *Trends in Food Science & Technology*, *20*(11–12), 557–566.

Chetwynd, A. J., & David, A. (2018). A review of nanoscale LC-ESI for metabolomics and its potential to enhance the metabolome coverage. *Talanta*, *182*, 380–390.

Cotter, R. J. (1998). *Time-of-Flight Mass Spectrometry: Instrumentation and Applications in Biological Research*. American Chemical Society.

Danezis, G. P., Tsagkaris, A. S., Brusic, V., & Georgiou, C. A. (2016). Food authentication: state of the art and prospects. *Current Opinion in Food Science*, *10*, 22–31.

Dass, C. (2007). *Fundamentals of contemporary mass spectrometry*. Wiley-Interscience.

De Vijlder, T., Valkenborg, D., Lemière, F., Romijn, E. P., Laukens, K., & Cuyckens, F. (2017). A tutorial in small molecule identification via electrospray ionization-mass spectrometry: The practical art of structural elucidation. *Mass Spectrometry Reviews*.

Dettmer, K., Aronov, P. A., & Hammock, B. D. (2007). Mass spectrometry-based metabolomics. *Mass Spectrometry Reviews*, *26*(1), 51–78.

Díaz, R., Gallart-Ayala, H., Sancho, J. V, Nuñez, O., Zamora, T., Martins, C. P. B., … Checa, A. (2016). Told through the wine: A liquid chromatography-mass spectrometry interplatform comparison reveals the influence of the global approach on the final annotated metabolites in non-targeted metabolomics. *Journal of Chromatography. A*, *1433*, 90–7.

Díaz, R., Pozo, O. J., Sancho, J. V, & Hernández, F. (2014). Metabolomic approaches for orange origin discrimination by ultra-high performance liquid chromatography coupled to quadrupole time-of-flight mass spectrometry. *Food Chemistry*, *157*, 84–93.

Duarte, I. F., Rocha, C. M., & Gil, A. M. (2013). Metabolic profiling of biofluids: potential in lung cancer screening and diagnosis. *Expert Review of Molecular Diagnostics*, *13*(7), 737–48.

Dunn, W. B., & Ellis, D. I. (2005). Metabolomics: Current analytical platforms and methodologies. *TrAC - Trends in Analytical Chemistry*, *24*(4), 285–294.

Estensoro, I., Ballester-Lozano, G., Benedito-Palos, L., Grammes, F., Martos-Sitcha, J. A., Mydland, L.-T., … Pérez-Sánchez, J. (2016). Dietary Butyrate Helps to Restore the Intestinal Status of a Marine Teleost (Sparus aurata) Fed Extreme Diets Low in Fish Meal and Fish Oil. *PLOS ONE*, *11*(11), e0166564.

Fiehn, O. (2001). Combining Genomics, Metabolome Analysis, and Biochemical Modelling to Understand Metabolic Networks. *Comparative and Functional Genomics*, *2*(3), 155–168.

Gabelica, V., & Marklund, E. (2018). Fundamentals of ion mobility spectrometry. *Current Opinion in Chemical Biology*, *42*, 51–59.

Galeano Diaz, T., Durán Merás, I., Sánchez Casas, J., & Alexandre Franco, M. F. (2005). Characterization of virgin olive oils according to its triglycerides and sterols composition by chemometric methods. *Food Control*, *16*(4), 339–347.

Garrido-Delgado, R., Dobao-Prieto, M. D. M., Arce, L., & Valc??rcel, M. (2015). Determination of

volatile compounds by GC-IMS to assign the quality of virgin olive oil. *Food Chemistry*, *187*, 572–579.

Gygi, S. P., Rochon, Y., Franza, B. R., & Aebersold, R. (1999). Correlation between protein and mRNA abundance in yeast. *Molecular and Cellular Biology*, *19*(3), 1720–30. Retrieved from

Huang, S. S. Y., Benskin, J. P., Chandramouli, B., Butler, H., Helbing, C. C., & Cosgrove, J. R. (2016). Xenobiotics Produce Distinct Metabolomic Responses in Zebrafish Larvae (*Danio rerio*). *Environmental Science & Technology*, *50*(12), 6526–6535.

Kamour, R., Ammar, A., El-Attug, M., & Almog, T. (2013). Development of fused-core silica HPLC columns and their recent pharmaceutical and biological applications: A review. *International Journal of Pharmacy and Pharmaceutical Sciences*.

Kell, D. B. (2004). Metabolomics and systems biology: making sense of the soup. *Current Opinion in Microbiology*, *7*(3), 296–307.

Lavine, B., & Workman, J. (2008). Chemometrics. *Analytical Chemistry*, *80*(12), 4519–4531.

Li, C., Li, P., Tan, Y. M., Lam, S. H., Chan, E. C. Y., & Gong, Z. (2016). Metabolomic Characterizations of Liver Injury Caused by Acute Arsenic Toxicity in Zebrafish. *PLOS ONE*, *11*(3), e0151225.

Malkar, A., Devenport, N. A., Martin, H. J., Patel, P., Turner, M. A., Watson, P., … Creaser, C. S. (2013). Metabolic profiling of human saliva before and after induced physiological stress by ultra-high performance liquid chromatography-ion mobility-mass spectrometry. *Metabolomics*, *9*(6), 1192–1201.

Mancano, G., Mora-Ortiz, M., & Claus, S. P. (2018). Recent developments in nutrimetabolomics: from food characterisation to disease prevention. *Current Opinion in Food Science*, *22*, 145–152.

Monasterio, R. P., Olmo-García, L., Bajoub, A., Fernández-Gutiérrez, A., & Carrasco-Pancorbo, A. (2017). Phenolic Compounds Profiling of Virgin Olive Oils from Different Varieties Cultivated in Mendoza, Argentina, by Using Liquid Chromatography-Mass Spectrometry. *Journal of Agricultural and Food Chemistry*, *65*(37), 8184–8195.

Olmo-García, L., Bajoub, A., Monasterio, R. P., Fernández-Gutiérrez, A., & Carrasco-Pancorbo, A. (2017). Metabolic profiling approach to determine phenolic compounds of virgin olive oil by direct injection and liquid chromatography coupled to mass spectrometry. *Food Chemistry*, *231*, 374–385.

Piazzon, M. C., Calduch-Giner, J. A., Fouz, B., Estensoro, I., Simó-Mirabet, P., Puyalto, M., … Pérez-Sánchez, J. (2017). Under control: how a dietary additive can restore the gut microbiome and proteomic profile, and improve disease resilience in a marine teleostean fish fed vegetable diets. *Microbiome*, *5*(1), 164.

Pimentel, G., Burton, K. J., Vergères, G., & Dupont, D. (2018). The role of foodomics to understand the digestion/bioactivity relationship of food. *Current Opinion in Food Science*, *22*, 67–73.

Pinto, R. C., Trygg, J., & Gottfries, J. (2012, June 1). Advantages of orthogonal inspection in chemometrics. *Journal of Chemometrics*. John Wiley & Sons, Ltd.

Poole, C. F. (2003). The Essence of Chromatography. *The Essence of Chromatography*, 847–899.

Ramses F. J. Kemperman, †,‡, Peter L. Horvatovich, †, Berend Hoekman, †, Theo H. Reijmers, §, Frits A. J. Muskiet, ‡ and, & Rainer Bischoff*, †. (2006). Comparative Urine Analysis by Liquid Chromatography−Mass Spectrometry and Multivariate Statistics: Method Development, Evaluation, and Application to Proteinuria.

Raro, M., Ibáñez, M., Gil, R., Fabregat, A., Tudela, E., Deventer, K., … Pozo, Ó. J. (2015). Untargeted Metabolomics in Doping Control: Detection of New Markers of Testosterone Misuse by Ultrahigh Performance Liquid Chromatography Coupled to High-Resolution Mass Spectrometry. *Analytical Chemistry*, *87*(16), 8373–8380.

Raterink, R. J., Lindenburg, P. W., Vreeken, R. J., Ramautar, R., & Hankemeier, T. (2014, October 1). Recent developments in sample-pretreatment techniques for mass spectrometry-based metabolomics. *TrAC - Trends in Analytical Chemistry*. Elsevier.

Riedl, J., Esslinger, S., & Fauhl-Hassek, C. (2015). Review of validation and reporting of non-targeted fingerprinting approaches for food authentication. *Analytica Chimica Acta*, *885*, 17–32.

Rodionova, O. Y., Titova, A. V., & Pomerantsev, A. L. (2016, April 1). Discriminant analysis is an inappropriate method of authentication. *TrAC - Trends in Analytical Chemistry*. Elsevier.

Sales, C., Cervera, M. I., Gil, R., Portol??s, T., Pitarch, E., & Beltran, J. (2017). Quality classification of Spanish olive oils by untargeted gas chromatography coupled to hybrid quadrupole-time of flight mass spectrometry with atmospheric pressure chemical ionization and metabolomics-based statistical approach. *Food Chemistry*, *216*, 365–373.

Sánchez de Medina, V., Riachy, M. El, Priego-Capote, F., & Luque de Castro, M. D. (2013). Mass spectrometry to evaluate the effect of the ripening process on phenols of virgin olive oils. *European Journal of Lipid Science and Technology*, *115*(9), 1053–1061.

Schymanski, E. L., Jeon, J., Gulde, R., Fenner, K., Ruff, M., Singer, H. P., & Hollender, J. (2014, February 18). Identifying small molecules via high resolution mass spectrometry: Communicating confidence. *Environmental Science and Technology*. American Chemical Society.

Shen, G., Huang, Y., Dong, J., Wang, X., Cheng, K.-K., Feng, J., … Ye, J. (2018). Metabolic Effect of Dietary Taurine Supplementation on Nile Tilapia (*Oreochromis niloticus*) Evaluated by NMR-Based Metabolomics. *Journal of Agricultural and Food Chemistry*, *66*(1), 368–377.

Shen, Q., Dong, W., Yang, M., Li, L., Cheung, H. Y., & Zhang, Z. (2013). Lipidomic fingerprint of almonds (prunus dulcis L. cv nonpareil) using TiO2 nanoparticle based matrix solid-phase dispersion and MALDI-TOF/MS and its potential in geographical origin verification. *Journal of Agricultural and Food Chemistry*, *61*(32), 7739–7748.

Simó-Mirabet, P., Perera, E., Calduch-Giner, J. A., Afonso, J. M., & Pérez-Sánchez, J. (2018). Co-expression Analysis of Sirtuins and Related Metabolic Biomarkers in Juveniles of Gilthead Sea Bream (Sparus aurata) With Differences in Growth Performance. *Frontiers in Physiology*, *9*, 608.

Smith, C. A., Want, E. J., O'Maille, G., Abagyan, R., & Siuzdak, G. (2006). XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal Chem*, *78*(3), 779–787.

Vrijheid, M. (2014). The exposome: a new paradigm to study the impact of environment on health. *Thorax*, *69*(9), 876–8.

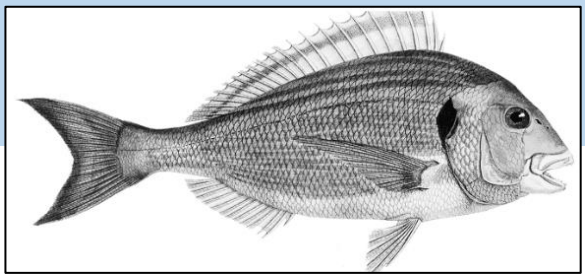
Wang, M., Rang, O., Liu, F., Xia, W., Li, Y., Zhang, Y., … Xu, S. (2018). A systematic review of metabolomics biomarkers for Bisphenol A exposure. *Metabolomics*, *14*(4), 45.

Whitehouse, C. M., Dreyer, R. N., Yamashita, M., & Fenn, J. B. (1985). Electrospray interface for liquid chromatographs and mass spectrometers. *Analytical Chemistry*, *57*(3), 675–679.

Wishart, D. S. (2009). Computational strategies for metabolite identification in metabolomics. *Bioanalysis*, *1*(9), 1579–1596.

Zontov, Y. V., Rodionova, O. Y., Kucheryavskiy, S. V., & Pomerantsev, A. L. (2017). DD-SIMCA – A MATLAB GUI tool for data driven SIMCA approach. *Chemometrics and Intelligent Laboratory Systems*, *167*, 23–28.

CAPÍTULO II



ESTUDIO DE DIFERENTES ESTADOS DE NUTRICIÓN Y COMPOSICIÓN DE DIETAS EN DORADAS

II.1: Artículo científico 1

PeerJ

Untargeted metabolomics approach for unraveling robust biomarkers of nutritional status in fasted gilthead sea bream (*Sparus aurata*)

Ruben Gil-Solsona[1], Jaime Nácher-Mestre[1,2], Leticia Lacalle-Bergeron[1], Juan Vicente Sancho[1], Josep Alvar Calduch-Giner[2], Félix Hernández[1] and Jaume Pérez-Sánchez[2]

[1] Research Institute for Pesticides and Water (IUPA), University Jaume I, Castellón, Spain
[2] Institute of Aquaculture Torre de la Sal (IATS, CSIC), Ribera de Cabanes, Castellón, Spain

ABSTRACT

A metabolomic study has been performed to identify sensitive and robust biomarkers of malnutrition in farmed fish, using gilthead sea bream (*Sparus aurata*) as a model. The metabolomic fingerprinting of serum from fasted fish was assessed by means of ultra-high performance liquid chromatography coupled to quadrupole time-of-flight mass spectrometry. More than 15,000 different *m/z* ions were detected and Partial Least Squares–Discriminant analysis allowed a clear differentiation between the two experimental groups (fed and 10-day fasted fish) with more than 90% of total variance explained by the two first components. The most significant metabolites (up to 45) were elucidated on the basis of their tandem mass spectra with a broad representation of amino acids, oligopeptides, urea cycle metabolites, L-carnitine-related metabolites, glutathione-related metabolites, fatty acids, lysophosphatidic acids, phosphatidylcholines as well as biotin- and noradrenaline-related metabolites. This untargeted approach highlighted important adaptive responses in energy and oxidative metabolism, contributing to identify robust and nutritionally-regulated biomarkers of health and metabolic condition that will serve to assess the welfare status of farmed fish.

Submitted 10 November 2016
Accepted 17 December 2016
Published 26 January 2017

Corresponding authors
Juan Vicente Sancho, sanchoj@uji.es
Jaume Pérez-Sánchez, jaime.perez.sanchez@csic.es

Academic editor
Kenneth Storey

Additional Information and Declarations can be found on page 14

DOI 10.7717/peerj.2920

Subjects Aquaculture, Fisheries and Fish Science, Biochemistry, Bioinformatics, Nutrition, Metabolic Sciences
Keywords Aquaculture, Nutrition, Chromatography, Mass spectrometry, Gilthead sea bream, Serum metabolomics

INTRODUCTION

Fish aquaculture is the sector of animal livestock production with higher growth rates at the global level. This industry highly contributes to cover the current but also the future demand of nutritious quality food for human consumption (*Ottinger, Clauss & Kuenzer, 2016*). This starts with the selection of high quality raw materials in order to ensure the development of an efficient and environmentally sustainable sector. However, we need to refine our knowledge on nutrient requirements to produce more robust, safe and quality fish, especially with the advent of new diet formulations based on alternative plant ingredients rather than marine feedstuffs (*Karalazos et al., 2007*; *Médale et al., 2013*; *Benedito-Palos et al.,*

How to cite this article Gil-Solsona et al. (2017), Untargeted metabolomics approach for unraveling robust biomarkers of nutritional status in fasted gilthead sea bream (*Sparus aurata*). PeerJ 5:e2920; DOI 10.7717/peerj.2920

Untargeted metabolomics approach for unraveling robust biomarkers of nutritional status in fasted gilthead sea bream (*Sparus aurata*)

Rubén Gil-Solsona[1], Jaime Nácher-Mestre[1,2], Leticia Lacalle-Bergeron[1], Juan Vicente Sancho[1], Josep Alvar Calduch-Giner[2], Félix Hernández[1], Jaume Pérez-Sánchez[2]

[1] Research Institute for Pesticides and Water (IUPA), University Jaume I, Castellón, Spain.

[2] Institute of Aquaculture Torre de la Sal (IATS, CSIC), Ribera de Cabanes, Castellón, Spain.

Abstract

A metabolomic study has been performed to identify sensitive and robust biomarkers of malnutrition in farmed fish, using gilthead sea bream (*Sparus aurata*) as a model. The metabolomic fingerprinting of serum from fasted fish was assessed by means of ultra-high performance liquid chromatography coupled to quadrupole time-of-flight mass spectrometry. More than 15,000 different *m/z* ions were detected and Partial Least Squares - Discriminant analysis allowed a clear differentiation between the two experimental groups (fed and 10-day fasted fish) with more than 90 % of total variance explained by the two first components. The most significant metabolites (up to 45) were elucidated on the basis of their tandem mass spectra with a broad representation of amino acids, oligopeptides, urea cycle metabolites, L-carnitine-related metabolites, glutathione-related metabolites, fatty acids, lysophosphatidic acids, phosphatidylcholines as well as biotin- and noradrenaline-related metabolites. This untargeted approach highlighted important adaptive responses in energy and oxidative metabolism, contributing to identify robust and nutritionally-regulated biomarkers of health and metabolic condition that will serve to assess the welfare status of farmed fish.

Introduction

Fish aquaculture is the sector of animal livestock production with higher growth rates at the global level. This industry highly contributes to cover the current but also the future demand of nutritious quality food for human consumption (*Ottinger, Clauss, & Kuenzer, 2016*). This starts with the selection of high quality raw materials in order to ensure the development of an efficient and environmentally sustainable sector. However, we need to refine our knowledge on nutrient

requirements to produce more robust, safe and quality fish, especially with the advent of new diet formulations based on alternative plant ingredients rather than marine feedstuffs (*Karalazos et al., 2007; Médale et al., 2013; Benedito-Palos et al., 2016*). As a result, research in fish nutrition is moving from classical methodologies to omics approaches, including transcriptomics (*Laura Benedito-Palos, Ballester-Lozano, & Pérez-Sánchez, 2014; Louro, Power, & Canario, 2014*), proteomics (*Rodrigues, Silva, Dias, & Jessen, 2012; Wrzesinski et al., 2013*) and metabolomics (*Kullgren et al., 2010; Silva et al., 2014; Asakura et al., 2014*).

Unlike nucleic acid or protein-based omic techniques, metabolomics has to deal with low-molecular weight metabolic entities (< 1,000 Da) with diverse chemical and physical properties (*Kell, 2004*), which can vary from millimolar to picomolar concentrations. Two are the main analytical platforms currently used in metabolomics studies: nuclear magnetic resonance (NMR) (*Emwas, 2015*) and mass spectrometry (MS) (*Castro-Puyana & Herrero, 2013*). Most of the studies of metabolomic profiling or fingerprinting of body fluids in livestock animals are based on NMR approaches due to its great robustness and elucidation power (*Jégou et al., 2015; Kullgren et al., 2010; Niu, Li, Du, & Qin, 2016; H.-D. Xu et al., 2015*), although one of the main drawbacks of this technique is its low sensitivity (*Emwas, 2015*). By contrast, MS analyzers coupled to gas chromatography (GC) or high-performance liquid chromatography (HPLC) offer a high sensitivity, becoming a highly feasible and informative technique that has demonstrated its potential in human metabolomic studies (*Castro-Puyana & Herrero, 2013; Xu et al., 2009*). Besides, both NMR- and MS-based metabolomics rely on wide-untargeted approaches, but MS also allows retrospective analysis of relevant metabolites by means of the full-spectra acquisition by quadrupole time-of-flight mass analyzer (QTOF). Taking in mind all these constraints and advantages, a major aim of this study was to demonstrate the validity of metabolomics based on ultra-high performance liquid chromatography (UHPLC) and high resolution MS (HRMS) to provide new insights on the nutritional and metabolic phenotyping of farmed fish. To this end, the present work was conceived as a MS approach to identify and, most importantly, validate robust biomarkers of malnutrition in short-term fasted fish, using gilthead sea bream (*Sparus aurata*) as a model of a highly cultured fish in all the Mediterranean basin.

Materials & methods

Reagents and chemicals

HPLC-grade water was obtained from a Mili-Q water purification system (Millipore Ltd., Bedford, MA, USA). HPLC-grade methanol (MeOH), HPLC-supergradient acetonitrile (ACN), sodium hydroxide (> 99 %) and reagent grade ammonium acetate (NH_4Ac) were obtained from Scharlab (Barcelona, Spain). Leucine-enkephalin (mass-axis calibration), formic acid (mobile phase modifier) and analytical-grade standards methionine sulfoxide and trimethylamine N-oxide were purchased from Sigma-Aldrich (Saint Louis, MO, USA).

Animal care and sampling

Two-year-old gilthead sea bream of Atlantic origin (average initial weight: 380 g) were reared from early life stages in the indoor experimental facilities of the Institute of Aquaculture Torre de la Sal (IATS), following natural light and temperature conditions at our latitude (40º5'N, 0º10'E). The oxygen content of water was always higher than 85 % saturation, unionized ammonia remained below toxic levels (<0.02 mg/l), and rearing density was maintained lower than 15 kg/m3.

At mid-summer (July 2014), 30 fish were randomly allocated in two tanks (500 L). One group continued to be fed with a standard commercial diet (Biomar, EFICO Forte 824) to visual satiety one time per day, whereas the other group remained unfed for a 10-day period. At the end of this period, 10 fish from fasted and fed groups (following overnight fasting) were randomly sampled and anaesthetized with 100 mg/L of aminobenzoic acid ethyl ester (MS-222, Sigma-Aldrich) for blood and tissue sampling. Blood was taken from caudal vessels with vacutainer tubes with a clot activator. Liver and visceral adipose tissue were extracted and weighed. Blood samples were allowed to clot for 30 min at room temperature, and then centrifuged at 1,300 g for 10 min. The obtained samples were stored at -20 ºC until analysis.

All procedures were approved by the IATS Ethics and Animal Welfare Committee according to national (Royal Decree RD53/2013) and EU legislation (2010/63/EU) on the handling of animals for experiments.

Sample processing

Serum samples were centrifuged at 12,500 g for 10 min. Supernatant (400 µL) was diluted with 1.2 mL of ACN followed by centrifugation (12,500 g for 10 min). Then, 750 µL of supernatant were stored for hydrophilic interaction liquid chromatography (HILIC), and another 750 µL aliquot was evaporated to dryness by MiVac Duo Concentrator (40 ºC, 60 min) and dissolved with MeOH (75 µL) and Mili-Q Water (675 µL) for reversed phase (RP) analysis (details in **Fig. S1**). Quality control (QC) samples were prepared by pooling 50 µL of each sample extract. All samples were stored at -20 ºC until injection.

UHPLC-HRMS

A Waters Acquity UPLC system (Waters, Milford, MA, USA) was coupled to a hybrid quadrupole-TOF mass spectrometer (Xevo G2 QTOF, Waters, Manchester, UK), using a Z-spray-ESI interface operating in positive and negative ionization mode. The UHPLC separation was performed using Acquity UPLC® BEH C18 1.7 µm particle size analytical column 100 x 2.1 mm (Waters) at 300 µL/min flow rate for RP analysis. An Acquity UPLC® HILIC 1.7 µm particle size analytical column 100 x 2.1 mm (Waters) at 300 µL/min flow rate was used for hydrophilic interaction phase separations.

Each serum sample was injected four times, depending on the procedure (RP and HILIC) and the ionization mode selected (ESI+ and ESI-). The RP separation was performed using H2O with 0.01 % formic acid (HCOOH) as weak mobile phase (A) and MeOH with 0.01 % HCOOH as strong mobile phase (B). The percentage of B was changed from 10 % at 0 min, to 90 % at 14 min, 90% at 16 min and 10 % at 16.01 min, with a total run time of 18 min for both ESI+ and ESI-. For HILIC separation, the weak mobile phase was a mix of ACN:H_2O (95:5, v/v) with 0.01 % HCOOH and 10 mM NH_4Ac (A), and the strong mobile phase was H2O with 0.01 % HCOOH and 10 mM NH_4Ac (B). The B percentage was changed as follows: 0 min, 2 %; 1.5 min, 2 %; 2.5 min, 15 %; 6 min, 50 %; 7.5 min, 75 %; and finally at 7.51 min, 2 %, with a total run time of 10 min, for both ESI+ and ESI-. Sample injection volume was 10 µL in all cases. Nitrogen was used as both the desolvation gas and the nebulizing gas. A capillary voltage of 0.7 kV and 1.5 kV for positive and negative ion modes, respectively, and cone

voltage of 25 V were used. MS data were acquired over a *m/z* range of 50-1,200. TOF-MS resolution was approximately 20,000 at full width half maximum at *m/z* 556.2771. Collision gas was argon 99.995 % (Praxair, Valencia, Spain). The desolvation gas flow was set at 1,000 L/h, and the cone gas was set at 80 L/h. Desolvation gas temperature was set to 600 ℃, source temperature to 130 ℃, and column temperature to 40 ℃.

For MS^E experiments, two acquisition functions with different collision energies were created. The low energy (LE) function, with a fixed collision energy of 4 eV, and the high energy (HE) function, with a collision energy ramp ranging from 15 to 40 eV in order to obtain the (de)protonated ion from LE function and a wide range of fragment ions from the HE function. Both LE and HE functions used a scan time of 0.3 s with an inter-scan delay of 0.05 s. MS/MS experiments were carried out in the same conditions with different collision energies depending on the fragmentation observed for each compound. Calibrations were conducted from *m/z* 50 to 1,200 with a 1:1 mixture of 0.05 M NaOH:5 % HCOOH diluted (1:25) with H_2O:ACN (20:80), at a flow rate of 10 µL/min. For automated accurate mass measurement, a leucine-enkephalin solution (0.5 µg/mL) in ACN:H_2O (50:50) at 0.1 % HCOOH was pumped at 30 µL/min through the lock-spray needle and measured every 30 s, with a scan time of 0.3 s. The (de)protonated molecule of leucine-enkephalin, at *m/z* 556.2771 in positive mode and *m/z* 554.2615 in negative mode was used for recalibrating the mass axis during the injection and to ensure a robust accurate mass along time.

Data processing

The workflow of data processing is shown in **Fig. 1**. LC-MS spectral data were converted from proprietary (.raw, Waters Corp.) to generic (.cdf, NetCDF) format using Databridge application (within MassLynx v 4.1; Waters Corporation) and processed using XCMS R package (https://xcmsonline.scripps.edu/) (*Smith, Want, O'Maille, Abagyan, & Siuzdak, 2006*). Centwave feature detection algorithm was employed for peak picking (peak width from 5 to 20 s, S/N ratio higher than 10 and mass tolerance of 15 ppm) followed by retention time alignment for the detected features. Peak area normalization (mean centering) was applied to each data set in order to minimize instrumental drifts with a final log2 transformation to the area to standardize the range of independent feature variance followed by pareto scaling. ANOVA analysis followed by Benjamini-

Hochberg multiple testing correction was applied to the normalized peak areas of all metabolites to assess differences between fed and control groups.

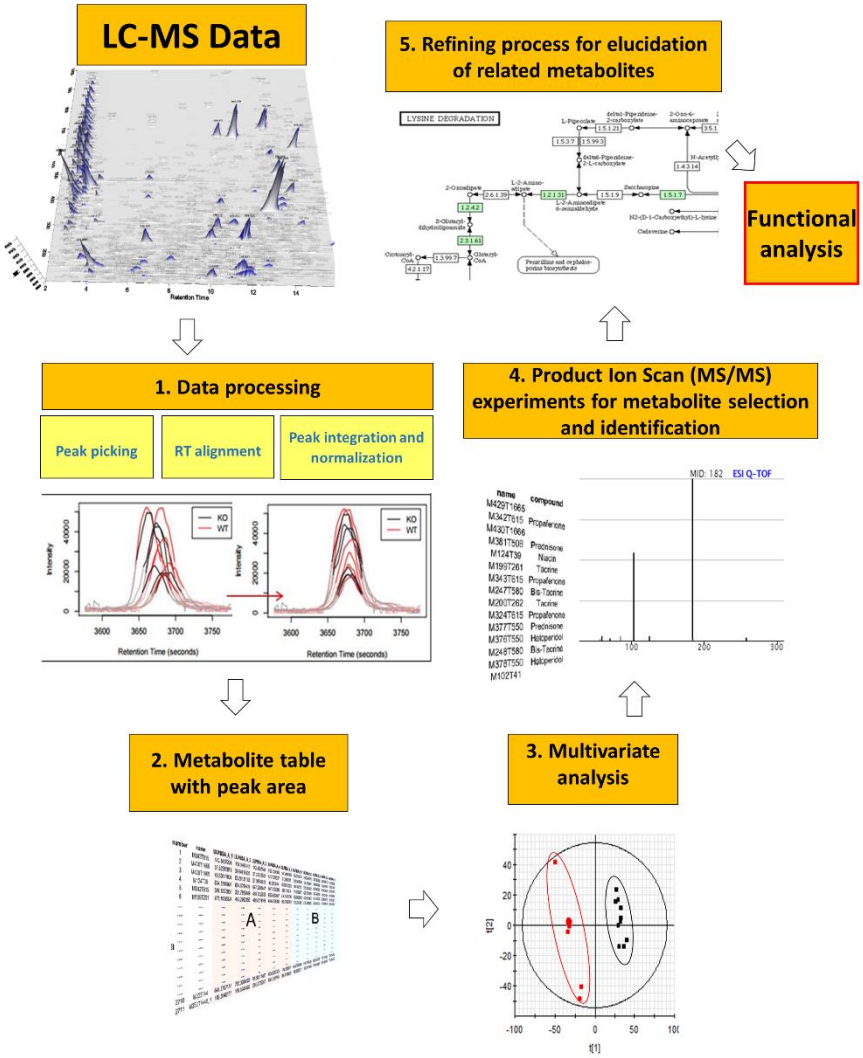


Fig. 1: General metabolomics workflow from data acquisition by LC-MS to functional analysis

Multivariate analysis of processed metabolomics data was performed by means of the EZ-Info software (Umetrics, Sweden). First, Principal Component Analysis (PCA) was employed to ensure the absence of outliers and the correct classification of QCs after normalization. Partial Least Squares

- Discriminant analysis (PLS-DA) was then applied to maximize the separation of fed and fasted individuals (*Fonville et al., 2010*). Orthogonal PLS-DA (OPLS-DA) was also carried out (*Wiklund et al., 2008*) with a high threshold (P[corr] > 0.95) for highlighting the most robust biomarkers.

For elucidation, the MS/MS spectra of the most significant metabolites were compared with reference spectra databases (METLIN, http://metlin.scripps.edu; Human Metabolome DataBase, http://www.hmbd.ca; MassBank, http://www.massbank.eu). For unassigned metabolites, in silico fragmentation software (MetFrag, http://msbi.ipb-halle.de/MetFrag) was employed, with subsequent searches through Chemspider (http://www.chemspider.com) and PubChem (https://pubchem.ncbi.nlm.nih.gov) chemical databases. Injection of standards of methionine sulfoxide and trimethylamine N-oxide served to validate the elucidation workflow.

A retrospective analysis of data previously acquired in MS^E mode served for the refined search of additional relevant metabolites. It consisted in the search of the *m/z* ratio (parent ions) of the metabolites of interest in the LE function as well as product ions obtained from MS/MS spectrum online databases (METLIN and Human Metabolome DataBase) in the HE function. Integrated areas of each candidate (parent ion) were compared in samples from fed and fasted groups.

Results and discussion

Biometric data

At the end of the experimental period, body weight of fed fish was 15 % higher than in fasted fish. This fasting protocol reduced the body fat depots, decreasing significantly ($P < 0.001$) the hepatosomatic index (100 x liver weight/body weight) from 1.3 to 0.9. The similar trend was found for mesenteric fat, although the decrease of mesenteric fat index (100 x mesenteric fat weight/body weight) from 1.9 to 1.6 was not statistically significant (**Table 1**). The magnitude of these changes was on the range of expected values for one- and two-year-old fish under similar experimental conditions (*Laura Benedito-Palos et al., 2014; Bermejo-Nogales, Calduch-Giner, & Pérez-Sánchez, 2015*).

Table 1. Biometry of fed (control) and fasted gilthead sea bream.

	Control	**Fasted**	**P-value**
Body weight (g)	426.5 ± 14.1	361.6 ± 10.7	0.002
Length (cm)	24.5 ± 0.3	24.0 ± 0.2	0.196
Condition factor	2.91 ± 0.06	2.63 ± 0.07	0.008
Liver weight (g)	5.63 ± 0.33	3.24 ± 0.16	$4E^{-6}$
Mesenteric fat (g)	7.20 ± 1.08	6.01 ± 0.89	0.408
HSI (%)[1]	1.32 ± 0.05	0.90 ± 0.03	$1E^{-6}$
MSI (%)[2]	1.9 ± 0.21	1.64 ± 0.23	0.394

[1]Hepatosomatic index = (100 x liver weight) / body weight
[2]Mesenteric fat index = (100 x mesenteric fat) / body weight

Untargeted metabolomics fingerprinting

Despite of the great potential of GC for chromatographic separation, the nature of serum samples, with medium-high polar compounds in a water-based fluid, pointed out to LC as a more convenient separation technique. UHPLC with sub-2 μm particle size was applied due to its high reproducibility and high separation performance in short-run time analyses. The use of different chromatographic techniques is a key issue to achieve a maximum of detected features when dealing with complex matrices like blood. In our case, serum samples were analyzed with two ionization modes and two different chromatographic columns: RP for a better separation of non-polar compounds, and HILIC to best separate the most polar compounds. In the RP analysis, 6,961 and 3,047 features were detected in both positive and negative ionization modes, while 4,820 and 1,015 features were labeled by XCMS using HILIC separation. This high total number of detected features (*m/z* values) highlights the huge detection power and sensitivity of HRMS and makes feasible a wide-view of sample composition to discriminate the most robust markers of nutritional conditions. Many features were only observed under a single ionization mode and chromatography type, reinforcing the importance of employing different chromatographic columns. As an example, a single peak was

detected by HILIC for the significant feature elucidated as LysoPC(20:5) while RP chromatography was able to separate ω-3 and ω-6 isomers (**Fig. S2**).

PLS-DA (of RP and HILIC in both positive and negative ionization modes) clearly discriminated the fasted individuals from those of the fed group (**Fig. 2**). Both groups were separated along the first component of the analysis, which explained 85-97 % of the total variance. Individuals of the same group were distributed along the second PLS-DA component (2-10 % of total variance). In the case of OPLS-DA, around 850 features from all four datasets were highlighted as discriminatory between fed and fasted fish with a P[corr] > 0.95 and a corrected P-value < 0.05 (see **Fig. S3**). Among them, up to 45 different compounds were elucidated as amino acids (4), oligopeptides (8), urea cycle-related metabolites (2), acylcarnitines (5), glutathione-related compounds (5), fatty acids (5), 3-hydroxyisovaleric acid, 3-methoxy-4-hydroxy-phenylglycol (MOPEG) sulphate and phospholipids (14), including phosphatidylcholines (PC) and lysoPC (**Table 2**).

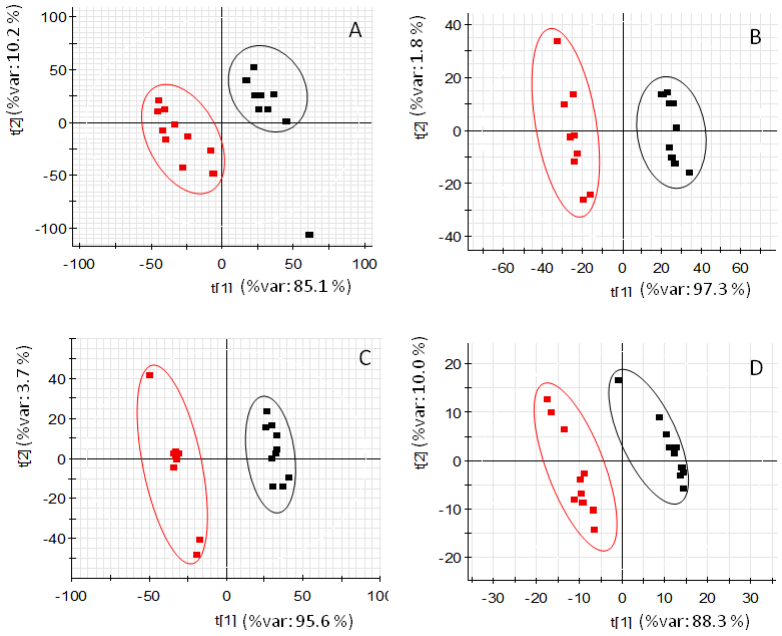


Fig. 2: PLS-DA score plots of acquired data of fasted (red) and control (black) fish. X-axis corresponds to first component and Y -axis to second component (A) RP at positive ionization mode (B) RP at negative ionization mode (C) HILIC at positive ionization mode (D) HILIC in negative ionization mode.

Table 2. Compound list obtained from untargeted approach and refining process.

	Compound Name	Biological process[a]	Chromatography / Ionization mode [b]	Formula	de/protonated molecule *m/z* (Error mDa)	RT (min)	Change (%) from CTRL[c]	Corrected P-value [d]
1	Octanoyl-L-carnitine	1,2	RP/+	$C_{15}H_{29}NO_4$	288.2158 (-1.8)	10.38	373 %	$3.1E^{-06}$
2	Decanoyl-L-carnitine	1,2	RP/+	$C_{17}H_{33}NO_4$	316.2478 (-1.0)	12.58	492 %	$4.53E^{-07}$
3	Hexadecenedioic acid mono-L-carnitine ester	1,2	RP/+	$C_{23}H_{43}NO_6$	430.3157 (-1.2)	13.28	373 %	$1.15E^{-06}$
4	Tetradecadien-L-carnitine	1,2	RP/+	$C_{21}H_{37}NO_4$	368.2784 (-1.7)	8.90	373 %	$1.25E^{-05}$
5	Tetradecenoyl-L-carnitine	1,2	RP/+	$C_{21}H_{39}NO_4$	370.2947 (-1.0)	14.88	373 %	$9.98E^{-06}$
6	L-ornithine	3,4	RP/+	$C_5H_{12}N_2O_2$	133.0972 (-0.5)	5.29	1213 %	$1.26E^{-10}$
7	Citrulline[e]	3,4	HI/+	$C_6H_{13}N_3O_3$	176.1029 (-0.6)	5.30	140 %	$5.02E^{-03}$
8	Argininosuccinate[e]	3,4	HI/-	$C_{10}H_{18}N_4O_6$	289.1145 (-0.3)	1.31	248 %	$2.07E^{-02}$
9	Norvaline	3,4	RP/-	$C_5H_{11}NO_2$	116.0711 (-0.1)	2.51	29 %	$5.54E^{-08}$
10	L- Arginine[e]	3,4	HI/+	$C_6H_{14}N_4O_2$	175.1194 (-0.1)	6.14	128 %	$3.37E^{-02}$
11	N-(2-cyanoethyl)glycine	3	HI/+	$C_5H_8N_2O_2$	129.0673 (+0.9)	5.25	246 %	$6.05E^{-09}$
12	Isoleucine	3	RP/-	$C_6H_{13}NO_2$	130.0880 (+1.4)	1.61	230 %	$1.15E^{-06}$
13	Glutamine	3	HI/+	$C_5H_{10}N_2O_3$	147.0766 (-0.4)	5.15	214 %	$1.80E^{-05}$
14	Glu-Phe	3	RP/+	$C_{14}H_{18}N_2O_5$	295.1278 (-1.6)	4.61	23 %	$9.11E^{-07}$
15	His-Phe	3	RP/+	$C_{15}H_{18}N_4O_3$	303.1447 (-1.0)	2.08	50 %	$6.53E^{-05}$
16	LTYV	3	RP/-	$C_{24}H_{38}N_4O_7$	493.2652 (-1.0)	5.80	23 %	$9.98E^{-06}$
17	LLGGPS	3	RP/+	$C_{24}H_{42}N_6O_8$	543.3148 (+0.6)	6.48	214 %	$2.53E^{-05}$
18	QLWD	3	RP/+	$C_{26}H_{36}N_6O_8$	561.2688 (+1.5)	8.08	373 %	$4.58E^{-04}$
19	YLWV	3	RP/+	$C_{24}H_{38}N_4O_7$	495.2807 (-1.2)	8.10	283 %	$2.77E^{-04}$
20	SVLGPA	3	RP/+	$C_{24}H_{42}N_6O_8$	543.3141 (-0.1)	14.75	746 %	$2.87E^{-07}$
21	N(2-Furoyl)glycyl-leucine	3	RP/-	$C_{13}H_{20}N_2O_5$	283.1301 (+0.7)	5.15	303 %	$3.44E^{-09}$
22	MOPEG sulphate	5	RP/-	$C_9H_{12}O_7S$	263.0223 (-0.2)	2.01	325 %	$9.70E^{-07}$
23	Noradrenaline[e]	5	HI/+	$C_8H_{11}NO_3$	170.0806 (-1.1)	3.02	141 %	$9.34E^{-03}$
24	DOPEGAL [e]	5	HI/+	$C_8H_8O_4$	169.0501 (0.0)	3.11	229 %	$3.94E^{-07}$
25	DOPEG [e]	5	HI/+	$C_8H_{10}O_4$	171.0649 (-0.8)	2.55	108 %	$1.94E^{-01}$
26	Y-Glu-Leu	6	HI/+	$C_{11}H_{20}N_2O_5$	261.1432 (-1.8)	3.37	246 %	$3.13E^{-06}$
27	Y-Glu-Val	6	RP/+	$C_{10}H_{18}N_2O_5$	247.1285 (-0.9)	2.15	246 %	$1.05E^{-08}$
28	Y-Glu-Ile	6	HI/+	$C_{11}H_{20}N_2O_5$	261.1446 (-0.4)	3.12	696 %	$1.48E^{-08}$
29	Pyroglutamic acid	6	HI/+	$C_5H_7NO_3$	130.0521 (-1.7)	5.15	303 %	$4.26E^{-06}$
30	Glutathione[e]	6	HI/-	$C_{10}H_{17}N_3O_6S$	306.0760 (0.0)	1.37	66 %	$3.03E^{-02}$
31	γ-Glu-Cys[e]	6	HI/+	$C_8H_{14}N_2O_5S$	251.0701 (-0.1)	4.92	63 %	$4.22E^{-01}$
32	Methionine sulfoxide	6	HI/+	$C_5H_{11}NO_3S$	166.0511 (-0.2)	4.96	41 %	$2.08E^{-04}$
33	FFA (C18:3)	1,2	RP/-	$C_{18}H_{30}O_2$	277.2172 (-0.4)	15.41	246 %	$2.08E^{-04}$
34	9-hydroxy-octadecanoic acid	1,2	RP/-	$C_{18}H_{36}O_3$	299.2598 (+1.3)	14.98	528 %	$2.23E^{-07}$
35	Linoleic acid	1,2	RP/-	$C_{20}H_{34}O_2$	305.2500 (+1.9)	16.31	19 %	$3.43E^{-06}$
36	Eicosapentaenoic acid	1,2	RP/-	$C_{20}H_{30}O_2$	301.2152 (-1.6)	15.10	33 %	$2.01E^{-03}$
37	Acetohexadecyloxypropylaminoethylphosphate	1,7	RP/+	$C_{23}H_{48}NO_7P$	482.3250 (+0.3)	14.93	35 %	$2.53E^{-06}$
38	LysoPC(14:0)	1,7	RP/+	$C_{22}H_{46}NO_7P$	468.3083 (-0.7)	15.41	25 %	$9.29E^{-05}$
39	LysoPC(16:1)	1,7	RP/+	$C_{24}H_{48}NO_7P$	494.3244 (-0.3)	14.91	44 %	$1.71E^{-04}$
40	LysoPC(18:4)	1,7	RP/+	$C_{26}H_{46}NO_7P$	516.3079 (-1.1)	14.36	11 %	$1.40E^{-08}$
41	LysoPC(18:3)	1,7	RP/+	$C_{26}H_{48}NO_7P$	518.3247 (0.0)	14.38	4 %	$1.93E^{-11}$
42	LysoPC(20:5)	1,7	RP/+	$C_{28}H_{48}NO_7P$	542.3248 (+0.1)	14.58	13 %	$1.37E^{-08}$
43	LysoPC(18:1)	1,7	RP/+	$C_{26}H_{52}NO_7P$	522.3560 (0.0)	15.00	16 %	$8.10E^{-08}$
44	LysoPC(21:5)	1,7	RP/+	$C_{29}H_{50}NO_7P$	556.3408 (+0.5)	14.88	31 %	$9.98E^{-06}$
45	LysoPC (20:4)	1,7	RP/+	$C_{28}H_{50}NO_7P$	544.3411 (+0.8)	14.06	20 %	$7.01E^{-06}$
46	LysoPC(Tetradecylthioacetic acid)	1,7	RP/+	$C_{24}H_{48}NO_7PS$	526.2951 (-1.6)	15.13	1 %	$1.62E^{-09}$
47	PC(22:5/20:5)	1,7	HI/+	$C_{50}H_{80}NO_8P$	854.5689 (-1.1)	4.11	27 %	$3.01E^{-08}$
48	PC(20:5/18:1)	1,7	HI/+	$C_{46}H_{78}NO_8P$	804.5435 (-0.8)	6.78	200 %	$1.95E^{-02}$
49	PC(20:4/18:1)	1,7	HI/+	$C_{46}H_{82}NO_8P$	808.5843 (-1.3)	5.73	214 %	$4.69E^{-02}$
50	PC(18:1/16:0)	1,7	HI/+	$C_{42}H_{82}NO_8P$	760.5847 (-0.9)	4.06	303 %	$1.79E^{-05}$
51	TMAO	1,7	HI/+	C_3H_9NO	76.0760 (-0.2)	5.85	41 %	$7.68E^{-05}$
52	Glycerophosphocholine	1,7	HI/+	$C_8H_{20}NO_6P$	258.1106 (0.0)	5.73	12 %	$5.69E^{-05}$
53	Hydroxyisovaleric acid	8	RP/-	$C_5H_{10}O_3$	117.0553 (+0.1)	2.85	373 %	$2.41E^{-09}$
54	Biotin[e]	8	RP/+	$C_{10}H_{16}N_2O_3S$	245.0959 (-0.1)	5.40	80 %	$2.46E^{-01}$

a. 1, Lipid metabolism; 2, Fatty acid metabolism; 3, Protein metabolism; 4, Amino catabolism/urea-cycle; 5 Stress response/catecholamine metabolism; 6, Oxidative stress/glutathione metabolism; 7, Phospholipid metabolism; 8, Vitamin metabolism.

b. Chromatography and ionization modes in which the signal area was higher for the highlighted compound.

c. Variation of area between fed and fasted fish. Variation > 100 % means higher area in fasted fish, and < 100 % means lower are in fasted fish.

d. ANOVA followed by Benjamini-Hochberg multiple testing correction.

e. Compounds obtained in refining process.

Phospholipids were characterized by the presence of both the protonated molecule and sodium adduct in the positive LE spectra and their acetate adducts in negative LE spectra. As an example, **Fig. S4** shows the *m/z* 542.3248 (+0.1 mDa mass error) which corresponds to the protonated molecule, and *m/z* 564.3074 (+1.2 mDa mass error) corresponds to the sodium adduct in RP+. MS/MS experiments were also carried out obtaining *m/z* 184.0735 (-0.4 mDa) (main product ions at 30 eV) in positive ionization mode. These *m/z* ions were annotated as phosphocholine in agreement with fragmentation pathways of lysoPC (*González-Domínguez, García-Barrera, & Gómez-Ariza, 2014; F. Xu, Zou, Lin, & Ong, 2009*).

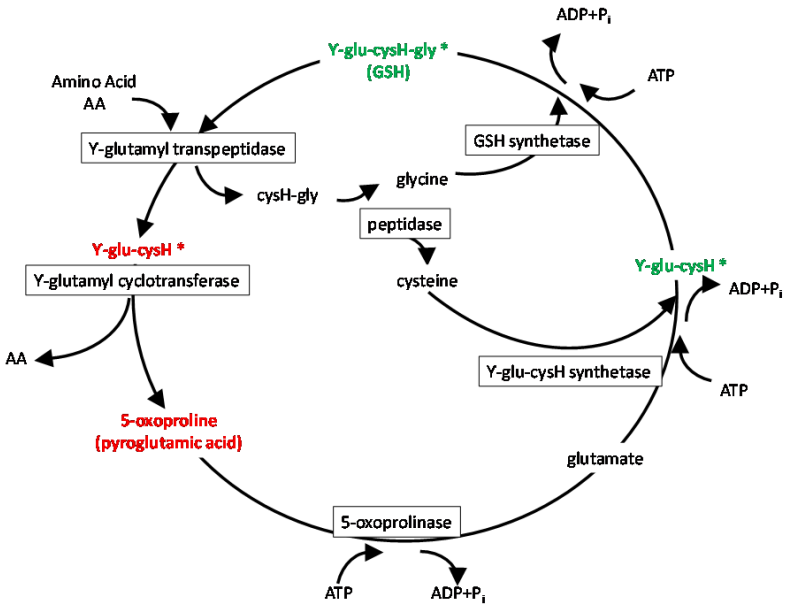


Fig. 3: Meister's cycle. In red, elucidated metabolites up-regulated with fasting; in green, downregulated with fasting. Asterisks mark elucidated metabolites by means of the refining step.

Carnitine-related compounds were also elucidated after MS/MS experiments by the presence of *m/z* 144.1051 ($C_7H_{13}NO_2^+$, +2.6 mDa), 85.0303 ($C_4H_5O_2^+$, +1.3 mDa) and 60.0814 ($C_3H_{10}N^+$, -0.3 mDa) as characteristic product ions of these compounds (*Luci, Hirche, & Eder, 2008; Möder, Löster, Herzschuh, & Popp, 1997*) also observed in METLIN spectra for some of these compounds (**Fig. S5**).

Tandem mass spectrometry also provides relevant information for isomers differentiation. For example, γ-Glu-Ile and γ-Glu-Leu presented the same molecular formula and close retention times (**Fig. S6**). MS/MS experiments revealed very similar spectra at 10 and 20 eV with the exception of the *m/z* 142.0499 which only appeared for the peak at 3.37 min. After comparing both spectra with METLIN, only γ-Glu-Leu spectrum showed this *m/z* ion at 20 eV. The formation of $C_4H_8^+$ have been observed and described in the literature much higher in isoleucine than in leucine (Squire, Beranová, & Wesdemiotis, 1995), so in leucine spectra the neutral loss of C4H8 can be observed while it does not appears in isoleucine spectra. This strategy was followed for the rest of elucidated compounds.

Functional analysis of elucidated compounds

Biological significance of the concurrent up- or down-regulation of most of the elucidated metabolites during fasting clearly stated that food deprivation increased mobilization of body energy stores and improved the oxidative capacity of metabolic fuels, which paralleled the onset of specific changes in the cell redox-balance. In this regard, the increased mobilization of body fat stores in fasted individuals, exemplified by the loss of liver and adipose tissue mass, was linked to the consistent increase of circulating levels of five sub-products of L-carnitine (compounds 1-5 in **Table 2**), a carrier of fatty acids across the inner mitochondrial membrane for their subsequent beta-oxidation (*Luci, Hirche & Eder, 2008; Ball, Urschel & Pencharz, 2007*). At the molecular level, this was early substantiated in similar experimental conditions by a marked up-regulated expression of the two carnitine palmitoyltransferase variants (CPT1A, CPT1B) of the skeletal muscle of gilthead sea bream (*Benedito-Palos, Ballester-Lozano & Pérez-Sánchez, 2014*), which was encompassed by the increased expression of a high representation (25 enzyme subunits) of regulatory and assembly factors of the five enzyme complex units (Complex I-V) of the mitochondrial respiratory chain (*Bermejo-Nogales et al., 2015*). Microarray gene expression profiling of either glycolytic or aerobic muscle tissues of fish fed to maintenance ration also indicates that nutrient scarcity is by itself a major factor driving switches in muscle protein turnover and mitochondrial activity (*Calduch-Giner et al., 2014*). In the present study, this was reinforced by the consistent fasting increase of serum concentrations of urea cycle-related compounds (citrulline, ornithine, argininosuccinate and arginine). Of note, the activity

of urea cycle enzymes is typically higher in carnivorous fish than in herbivorous and omnivorous fish species (*Chiu, Austic & Rumsey, 1986*), and our results highlighted that acyl-carnitine and urea cycle metabolites are specially sensitive to fasting-mediated changes in fatty acid and amino acid catabolism during negative energy balance.

Catecholamines are mobilized into fish circulation during a variety of stressful situations which require modulation of cardiorespiratory function or mobilization of energy reserves. The magnitude of change is dependent on the species and the type and intensity of stress imposed, although a wide range of stressors including hypoxia, hypercapnia, exhaustive and violent exercise, air exposure or anemia are considered strong activators of the hypothalamic-pituitary-interrenal (HPI) axis in fish (Reid, Bernier, & Perry, 1998). This also applies to short-term fasting (*De Pedro, Delgado, Gancedo, & Alonso-Bedate, 2003*), and the observed increase of MOPEG sulphate, a metabolite of norepinephrine degradation, can be viewed as part of the adaptive response of the HPI axis to cope with fasting hypoglycemia through the activation of lipolysis and gluconeogenesis. This notion was supported by the increased levels of noradrenaline, and other catecholamine metabolites 3,4-dihydroxymandelaldehyde and 3,4-dihydroxyphenylethyleneglycol (DOPEGAL and DOPEG), detected in the refined search step of analysis.

The primary enzymatic antioxidant defense system of living organisms is the glutathione (GSH) redox system that reduces hydrogen peroxide and lipid hydroperoxides at the expense of oxidizing GSH to its disulfide form (GSSG). Once oxidized, GSH can be reduced back by glutathione reductase, using NADPH as an electron donor, and previous studies in gilthead seam bream indicate that either absolute GSH levels or the GSH/GSSG ratio are regulated by dietary oils, increasing the total plasma antioxidant capacity with the increased unsaturation index of dietary oils of marine origin (Saera-Vila et al., 2005). Likewise, total plasma antioxidant capacity is increased in hypoxic fish with a switch from oxidative phosphorylation (OXPHOS) to anaerobic glycolysis (*Bermejo-Nogales, Calduch-Giner, & Pérez-Sánchez, 2014*), which results in reduced mitochondria oxygen consumption and enhanced NADH production from glycolysis (*Frezza et al., 2011*). Importantly, extension of life span is related in mammals and birds to low antioxidant levels and low rates of generation of reactive oxygen species (ROS) (*Lykkesfeldt & Svendsen, 2007; Pamplona et al., 2008*). Experimental evidence in rats also indicates that intermittent fasting affects redox balance in a tissue selective manner (*Chausse, Vieira-*

Lara, Sanchez, Medeiros, & Kowaltowski, 2015), and our fish metabolomic study highlighted that the depletion of serum GSH during short-term fasting was closely related to changes in the Meister´s γ-glutamyl cycle with a key role in the recovery and delivery of cysteine in the body (*Griffith, Bridges, & Meister, 1978*). This was supported by high circulating concentrations of Y-Glu-(Leu/Val/Ile) and pyroglutamic acid in the serum of fasted gilthead sea bream, whereas both GSH and Y-Glu-Cys were depleted (**Fig. 3**). This represents a complex trade-off with a reduced risk of oxidative stress, also highlighted by the decreased concentration of methionine sulfoxide, an oxidized form of methionine that is highly correlated with the risk of oxidative stress (*Weissbach et al., 2002*). In parallel, other short oligopeptides were either increased or decreased in the serum of fasted fish. It remains to be established if they have a physiological significance or are subproducts of protein hydrolysis.

The fatty acid composition of triacylglycerols (TAG) usually clears a close resemblance to dietary lipids (*L Benedito-Palos, Navarro, Kaushik, & Pérez-Sánchez, 2010*), whereas that of phospholipids is highly regulated and influenced in fish by environmental factors, including temperature and osmolarity (*Ibarz et al., 2005; Los & Murata, 2004*). The specific effects of ration size have been addressed in gilthead sea bream, and maintenance ratio significantly increased the retention of arachidonic acid (ARA) and docosahexanoic acid (DHA) in muscle phospholipids, whereas the fatty acid composition of storage lipids remained almost unchanged (*Laura Benedito-Palos, Calduch-Giner, Ballester-Lozano, & Pérez-Sánchez, 2013*). This lean muscle phenotype was linked in the present study to low plasma levels of linoleic acid and eicosapentaenoic acid, the precursors of ARA and DHA, respectively. At the same time, fasting induced an overall decrease of circulating lysoPC and glycerophosphocholine, whereas the effect on phosphatydylcholines was more selective depending of the composition of fatty acids. In any case, phospholipid metabolism is becoming highly regulated by feed intake at either blood or tissue level. Thus, phospholipids of skeletal muscle would act as a reservoir of long chain poly-unsaturated fatty acids with an enhanced expression of lipoprotein lipase-like, a TAG lipase isoform exclusive of fish linage which is highly expressed in muscle tissues and specifically up-regulated by feed restriction (*Laura Benedito-Palos et al., 2013; Rimoldi, Benedito-Palos, Terova, & Pérez-Sánchez, 2015*). This in turn would mediate, at least in part, the changes in the blood composition of phospholipid-related metabolites. This is extensive to trimethylamine N-oxide (TMAO), and high TMAO and choline concentrations are associated in humans with diabetes and

advanced cardio-metabolic risk profile (*Obeid et al., 2016*). In agreement with this, the opposite pattern was found herein in fish under a negative energy balance, which reinforces the close metabolic association between interrelated pathways of phospholipid and oxidative metabolism. Recent metabolomics studies have highlighted a concurrent depletion of reduced GSH and glycerophosphocholine in the gills of another fish species, the golden grey mullet (Liza aurata), as a response to mercury toxicity (*Cappello et al., 2016a; Cappello et al., 2016b*). This finding also reinforces the view that the response to different challenges such as malnutrition or pollutants toxicity and susceptibility (*Kokushi et al., 2016; Wang et al., 2016*) can be assessed by the analysis of common responsive metabolites, opening the possibility for future screening of the fish general welfare status through selected biomarkers.

Lastly, major changes in vitamin status are related in our experimental model to biotin metabolism. In humans, the impairment of renal reclamation of biotin results in an elevated urine concentration of 3-hydroxyisovaleric acid (*Mock, Malik, Stumbo, Bishop, & Mock, 1997*). Accordingly, we found that the fasting increase of this metabolite was concurrent with a low biotin availability (highlighted by refining analysis), which reinforces the role of 3-hydroxyisovaleric acid as a biomarker of B7 vitamin deficiency in a wide range of vertebrate species, including fish.

Conclusions

This metabolomics study has been performed to help fish physiologists and nutritionists to identify highly sensitive and robust biomarkers of malnutrition from a large list of affected compounds (about 850 ions) (see **Fig. 4** as a corollary). The MS^E acquisition mode of the involved QTOF allowed to simultaneously recording both low and high collision energy mass spectra. In this sense, as full scan data is acquired, the possibility of refining steps after elucidation gives HRMS a strong advantage compared to NMR. Further studies are underway to determine the potential of this powerful methodological approach, alone or in combination with other omics approaches, for the discovery and validation of new biomarkers of nutritional status in a wide-range of physiological conditions arising with the advent of new fish feed formulations.

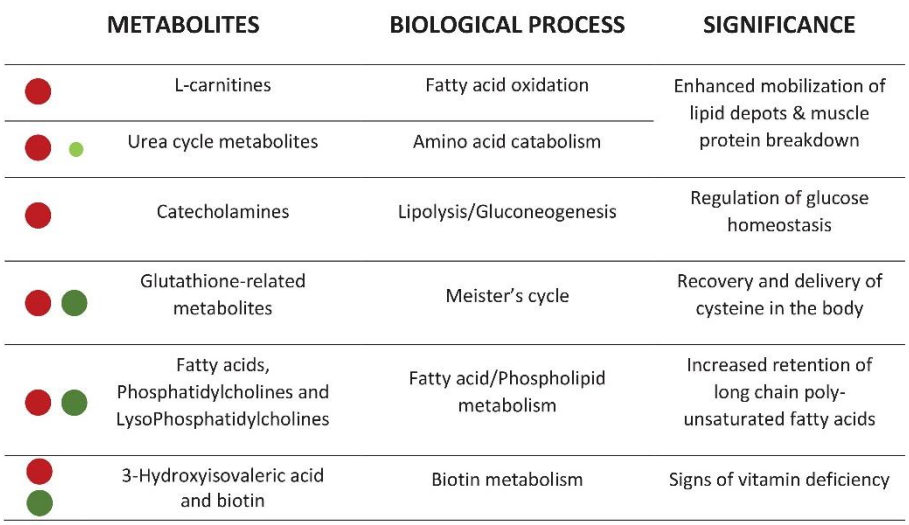


Fig. 4: Corollary with the metabolic significance of highlighted metabolites. Red and green circles signals the degree of up- and down-regulation of metabolites, respectively, with fasting.

Acknowledgments

This work has been developed in the framework of the Research Unit of Marine Ecotoxicology (IATS (CSIC)-IUPA (UJI)).

References

Asakura, T., Sakata, K., Yoshida, S., Date, Y., & Kikuchi, J. (2014). Noninvasive analysis of metabolic changes following nutrient input into diverse fish species, as investigated by metabolic and microbial profiling approaches. *PeerJ*, 2, e550.

Ball, R. O., Urschel, K. L., & Pencharz, P. B. (2007). Nutritional Consequences of Interspecies Differences in Arginine and Lysine Metabolism. *J. Nutr.*, 137(6), 1626S–1641.

Benedito-Palos, L., Ballester-Lozano, G. F., Simó, P., Karalazos, V., Ortiz, Á., Calduch-Giner, J., & Pérez-Sánchez, J. (2016). Lasting effects of butyrate and low FM/FO diets on growth performance, blood haematology/biochemistry and molecular growth-related markers in gilthead sea bream (*Sparus aurata*). *Aquaculture*, 454, 8–18.

Benedito-Palos, L., Ballester-Lozano, G., & Pérez-Sánchez, J. (2014). Wide-gene expression analysis of lipid-relevant genes in nutritionally challenged gilthead sea bream (*Sparus aurata*). *Gene*, 547(1), 34–42.

Benedito-Palos, L., Calduch-Giner, J. A., Ballester-Lozano, G. F., & Pérez-Sánchez, J. (2013). Effect of ration size on fillet fatty acid composition, phospholipid allostasis and mRNA expression patterns of lipid regulatory genes in gilthead sea bream (*Sparus aurata*). *The British Journal of Nutrition*, 109(7), 1175–87.

Benedito-Palos, L., Navarro, J. C., Kaushik, S., & Pérez-Sánchez, J. (2010). Tissue-specific robustness of fatty acid signatures in cultured gilthead sea bream (*Sparus aurata L.*) fed practical diets with a combined high replacement of fish meal and fish oil. *Journal of Animal Science*, 88(5), 1759–70.

Bermejo-Nogales, A., Calduch-Giner, J. A., & Pérez-Sánchez, J. (2014). Tissue-specific gene expression and functional regulation of uncoupling protein 2 (UCP2) by hypoxia and nutrient availability in gilthead sea bream (*Sparus aurata*): implications on the physiological significance of UCP1-3 variants. *Fish Physiology and Biochemistry*, 40(3), 751–62.

Bermejo-Nogales, A., Calduch-Giner, J. A., & Pérez-Sánchez, J. (2015). Unraveling the molecular signatures of oxidative phosphorylation to cope with the nutritionally changing metabolic capabilities of liver and muscle tissues in farmed fish. *PloS One*, 10(4), e0122889.

Calduch-Giner, J. A., Echasseriau, Y., Crespo, D., Baron, D., Planas, J. V, Prunet, P., & Pérez-Sánchez, J. (2014). Transcriptional assessment by microarray analysis and large-scale meta-analysis of the metabolic capacity of cardiac and skeletal muscle tissues to cope with reduced nutrient availability in Gilthead Sea Bream (*Sparus aurata L.*). *Marine Biotechnology* (New York, N.Y.), 16(4), 423–35.

Castro-Puyana, M., & Herrero, M. (2013). Metabolomics approaches based on mass spectrometry for food safety, quality and traceability. *TrAC Trends in Analytical Chemistry*, 52, 74–87.

Chausse, B., Vieira-Lara, M. A., Sanchez, A. B., Medeiros, M. H. G., & Kowaltowski, A. J. (2015). Intermittent Fasting Results in Tissue-Specific Changes in Bioenergetics and Redox State. *PLOS ONE*, 10(3), e0120413.

De Clercq, N., Bussche, J. Vanden, Van Meulebroek, L., Croubels, S., Delahaut, P., Buyst, D., Vanhaecke, L. (2015). Metabolic fingerprinting reveals a novel candidate biomarker for prednisolone treatment in cattle. *Metabolomics*, 12(1), 1.

De Pedro, N., Delgado, M. J., Gancedo, B., & Alonso-Bedate, M. (2003). Changes in glucose, glycogen, thyroid activity and hypothalamic catecholamines in tench by starvation and refeeding. *Journal of Comparative Physiology. B, Biochemical, Systemic, and Environmental Physiology*, 173(6), 475–81.

Emwas, A.-H. M. (2015). The strengths and weaknesses of NMR spectroscopy and mass spectrometry with particular focus on metabolomics research. *Methods in Molecular Biology* (Clifton, N.J.), 1277, 161–93.

Fonville, J. M., Richards, S. E., Barton, R. H., Boulange, C. L., Ebbels, T. M. D., Nicholson, J. K., … Dumas, M. E. (2010). The evolution of partial least squares models and related chemometric approaches in metabonomics and metabolic phenotyping. *Journal of Chemometrics*, 24(11–12), 636–649.

Frezza, C., Zheng, L., Tennant, D. A., Papkovsky, D. B., Hedley, B. A., Kalna, G., … Gottlieb, E. (2011). Metabolic profiling of hypoxic cells revealed a catabolic signature required for cell survival. *PloS One*, 6(9), e24411.

González-Domínguez, R., García-Barrera, T., & Gómez-Ariza, J. L. (2014). Combination of metabolomic and phospholipid-profiling approaches for the study of Alzheimer's disease. *Journal of Proteomics*, 104, 37–47.

Griffith, O. W., Bridges, R. J., & Meister, A. (1978). Evidence that the gamma-glutamyl cycle functions in vivo using intracellular glutathione: effects of amino acids and selective inhibition of enzymes. *Proceedings of the National Academy of Sciences*, 75(11), 5405–5408.

Ibarz, A., Blasco, J., Beltrán, M., Gallardo, M. A., Sánchez, J., Sala, R., & Fernández-Borràs, J. (2005). Cold-induced alterations on proximate composition and fatty acid profiles of several tissues in gilthead sea bream (*Sparus aurata*). *Aquaculture*, 249(1–4), 477–486.

Jégou, M., Gondret, F., Lalande-Martin, J., Tea, I., Baéza, E., & Louveau, I. (2015). NMR-based metabolomics highlights differences in plasma metabolites in pigs exhibiting diet-induced differences in adiposity. European Journal of Nutrition.

Karalazos, V., Bendiksen, E. Å., Dick,J. R., & Bell, J. G. (2007). Effects of dietary protein, and fat level and rapeseed oil on growth and tissue fatty acid composition and metabolism in Atlantic salmon (*Salmo salar L.*) reared at low water temperatures. *Aquaculture Nutrition*, 13(4), 256–265.

Kell, D. B. (2004). Metabolomics and systems biology: making sense of the soup. *Current Opinion in Microbiology*, 7(3), 296–307.

Kullgren, A., Samuelsson, L. M., Larsson, D. G. J., Björnsson, B. T., & Bergman, E. J. (2010). A metabolomics approach to elucidate effects of food deprivation in juvenile rainbow trout (*Oncorhynchus mykiss*). American Journal of Physiology. Regulatory, *Integrative and Comparative Physiology*, 299(6), R1440–R1448.

Los, D. A., & Murata, N. (2004). Membrane fluidity and its roles in the perception of environmental signals. *Biochimica et Biophysica Acta*, 1666(1–2), 142–57.

Louro, B., Power, D. M., & Canario, A. V. M. (2014). Advances in European sea bass genomics and future perspectives. *Marine Genomics*, 18, 71–75.

Luci, S., Hirche, F., & Eder, K. (2008). Fasting and caloric restriction increases mRNA concentrations of novel organic cation transporter-2 and carnitine concentrations in rat tissues. *Annals of Nutrition & Metabolism*, 52(1), 58–67.

Lykkesfeldt, J., & Svendsen, O. (2007). Oxidants and antioxidants in disease: oxidative stress in farm animals. *Veterinary Journal* (London, England : 1997), 173(3), 502–11.

Médale, F., Le Boucher, R., Dupont-Nivet, M., Quillet, E., Aubin, J., & Panserat, S. (2013). Des aliments à base de végétaux pour les poissons d'élevage. *Productions Animales*, 26(4), 303–316.

Mock, N. I., Malik, M. I., Stumbo, P. J., Bishop, W. P., & Mock, D. M. (1997). Increased urinary excretion of 3-hydroxyisovaleric acid and decreased urinary excretion of biotin are sensitive early indicators of decreased biotin status in experimental biotin deficiency. *American Journal of Clinical Nutrition*, 65, 951–958.

Möder, M., Löster, H., Herzschuh, R., & Popp, P. (1997). Determination of urinary acylcarnitines by ESI-MS coupled with solid-phase microextraction (SPME). *Journal of Mass Spectrometry* : *JMS*, 32(11), 1195–204.

Niu, Q.-Y., Li, Z.-Y., Du, G.-H., & Qin, X.-M. (2016). (1)H NMR based metabolomic profiling revealed doxorubicin-induced systematic alterations in a rat model. *Journal of Pharmaceutical and Biomedical Analysis*, 118, 338–48.

Obeid, R., Awwad, H. M., Rabagny, Y., Graeber, S., Herrmann, W., & Geisel, J. (2016). Plasma trimethylamine N-oxide concentration is associated with choline, phospholipids, and methyl metabolism. *The American Journal of Clinical Nutrition*, 103(3), 703–11.

Ottinger, M., Clauss, K., & Kuenzer, C. (2016). Aquaculture: Relevance, distribution, impacts and spatial assessments – A review. *Ocean & Coastal Management*, 119, 244–266.

Pamplona, R., Naudí, A., Gavín, R., Pastrana, M. A., Sajnani, G., Ilieva, E. V, … Requena, J. R. (2008). Increased oxidation, glycoxidation, and lipoxidation of brain proteins in prion disease. *Free Radical Biology & Medicine*, 45(8), 1159–66.

Putri, S. P., Nakayama, Y., Matsuda, F., Uchikata, T., Kobayashi, S., Matsubara, A., & Fukusaki, E. (2013). Current metabolomics: Practical applications. *Journal of Bioscience and Bioengineering*, 115(6), 579–589.

Reid, S. G., Bernier, N. J., & Perry, S. F. (1998). The adrenergic stress response in fish: control of catecholamine storage and release. *Comparative Biochemistry and Physiology Part C: Pharmacology, Toxicology and Endocrinology*, 120(1), 1–27.

Rimoldi, S., Benedito-Palos, L., Terova, G., & Pérez-Sánchez, J. (2015). Wide-targeted gene expression infers tissue-specific molecular signatures of lipid metabolism in fed and fasted fish. *Reviews in Fish Biology and Fisheries*, 26(1), 93–108.

Rodrigues, P. M., Silva, T. S., Dias, J., & Jessen, F. (2012). Proteomics in aquaculture: Applications and trends. *Journal of Proteomics*, 75(14), 4325–4345.

Saera-Vila, A., Calduch-Giner, J. A., Gómez-Requeni, P., Médale, F., Kaushik, S., & Pérez-Sánchez, J. (2005). Molecular characterization of gilthead sea bream (*Sparus aurata*) lipoprotein lipase. Transcriptional regulation by season and nutritional condition in skeletal muscle and fat storage tissues. *Comparative Biochemistry and Physiology. Part B, Biochemistry & Molecular Biology*, 142(2), 224–32.

Silva, T. S., da Costa, A. M. R., Conceição, L. E. C., Dias, J. P., Rodrigues, P. M. L., & Richard, N. (2014). Metabolic fingerprinting of gilthead seabream (*Sparus aurata*) liver to track interactions between dietary factors and seasonal temperature variations. *PeerJ*, 2, e527.

Smith, C. A., Want, E. J., O'Maille, G., Abagyan, R., & Siuzdak, G. (2006). XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal Chem*, 78(3), 779–787.

Squire, N. L., Beranová, Š., & Wesdemiotis, C. (1995). Tandem mass spectrometry of peptides. III—differentiation between leucine and isoleucine based on neutral losses. J*ournal of Mass Spectrometry*, 30(10), 1429–1434.

Weissbach, H., Etienne, F., Hoshi, T., Heinemann, S. H., Lowther, W. T., Matthews, B., … Brot, N. (2002). Peptide methionine sulfoxide reductase: structure, mechanism of action, and biological function. *Archives of Biochemistry and Biophysics*, 397(2), 172–178.

Wiklund, S., Johansson, E., Sjöström, L., Mellerowicz, E. J., Edlund, U., Shockcor, J. P., … Trygg, J. (2008). Visualization of GC/TOF-MS-based metabolomics data for identification of biochemically interesting compounds using OPLS class models. *Analytical Chemistry*, 80(1), 115–122.

Wrzesinski, K., R León, I., Kulej, K., Sprenger, R. R., Bjørndal, B., Christensen, B. J., … Rogowska-Wrzesinska, A. (2013). Proteomics identifies molecular networks affected by tetradecylthioacetic acid and fish oil supplemented diets. *Journal of Proteomics*, 84, 61–77.

Xu, F., Zou, L., Lin, Q., & Ong, C. N. (2009). Use of liquid chromatography/tandem mass spectrometry and online databases for identification of phosphocholines and lysophosphatidylcholines in human red blood cells. *Rapid Communications in Mass Spectrometry*, 23(19), 3243–3254.

Xu, H.-D., Wang, J.-S., Li, M.-H., Liu, Y., Chen, T., & Jia, A.-Q. (2015). $^{(1)}$H NMR based metabolomics approach to study the toxic effects of herbicide butachlor on goldfish (*Carassius auratus*). *Aquatic Toxicology* (Amsterdam, Netherlands), 159, 69–80.

Supplementary material

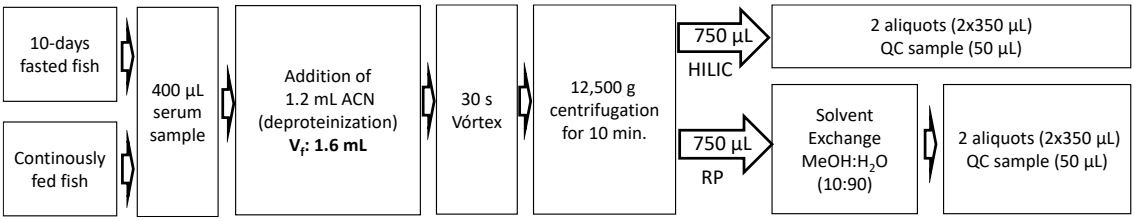


Fig. S1. Sample treatment for fish serum in both reversed phase and hydrophilic interaction liquid chromatography.

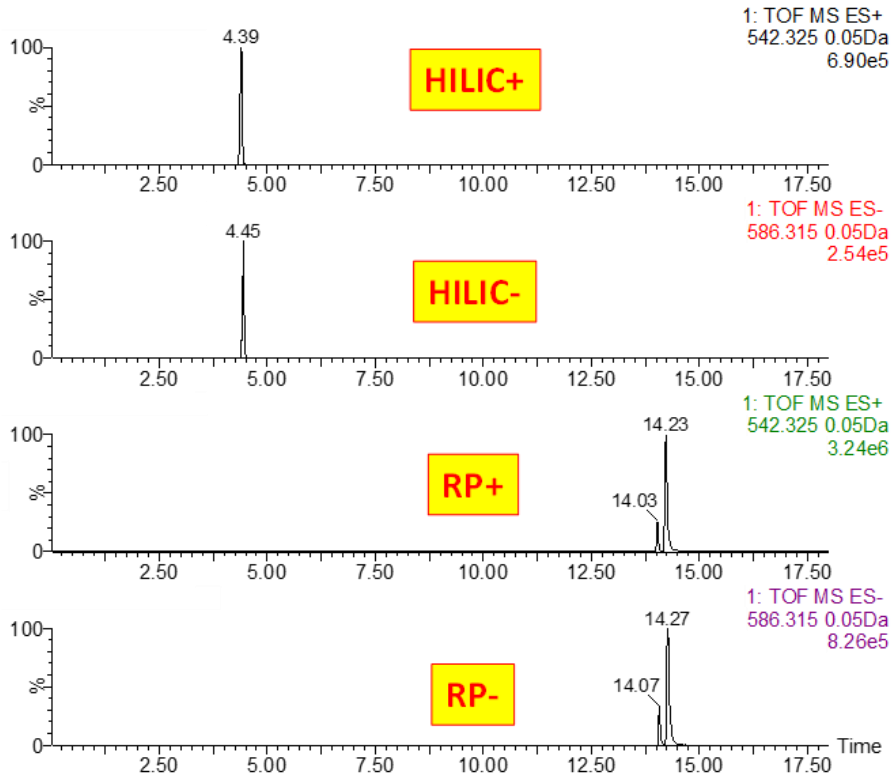


Fig. S2: A extracted ion chromatogram (XIC) for LysoPC(20:5) in HILIC and RP in both positive and negative ionization modes. In RP chromatography the separation of ω-3 and ω-6 LysoPC can be appreciated in both positive and negative ionization mode.

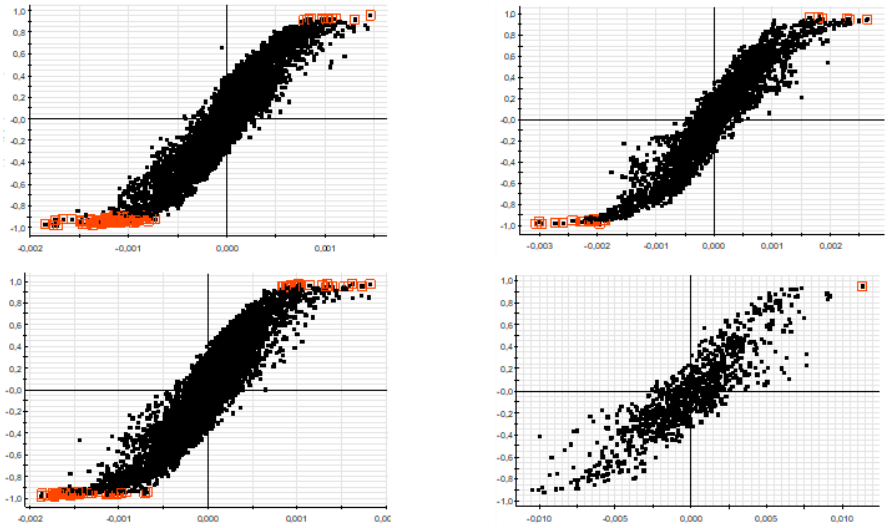


Fig. S3: Orthogonal PLS-DA S-Plot for individual biomarker highlighting. In red, ions with more than 0.95 correlation between groups. The ions enhanced by fasting are top-right and these ones decreased by fasting bottom-left on the plots. Top-left plot is the RP at positive ionization mode, top-right HILIC at positive ionization mode, bottom-left for RP at negative ionization mode and bottom-right HILIC in negative ionization mode.

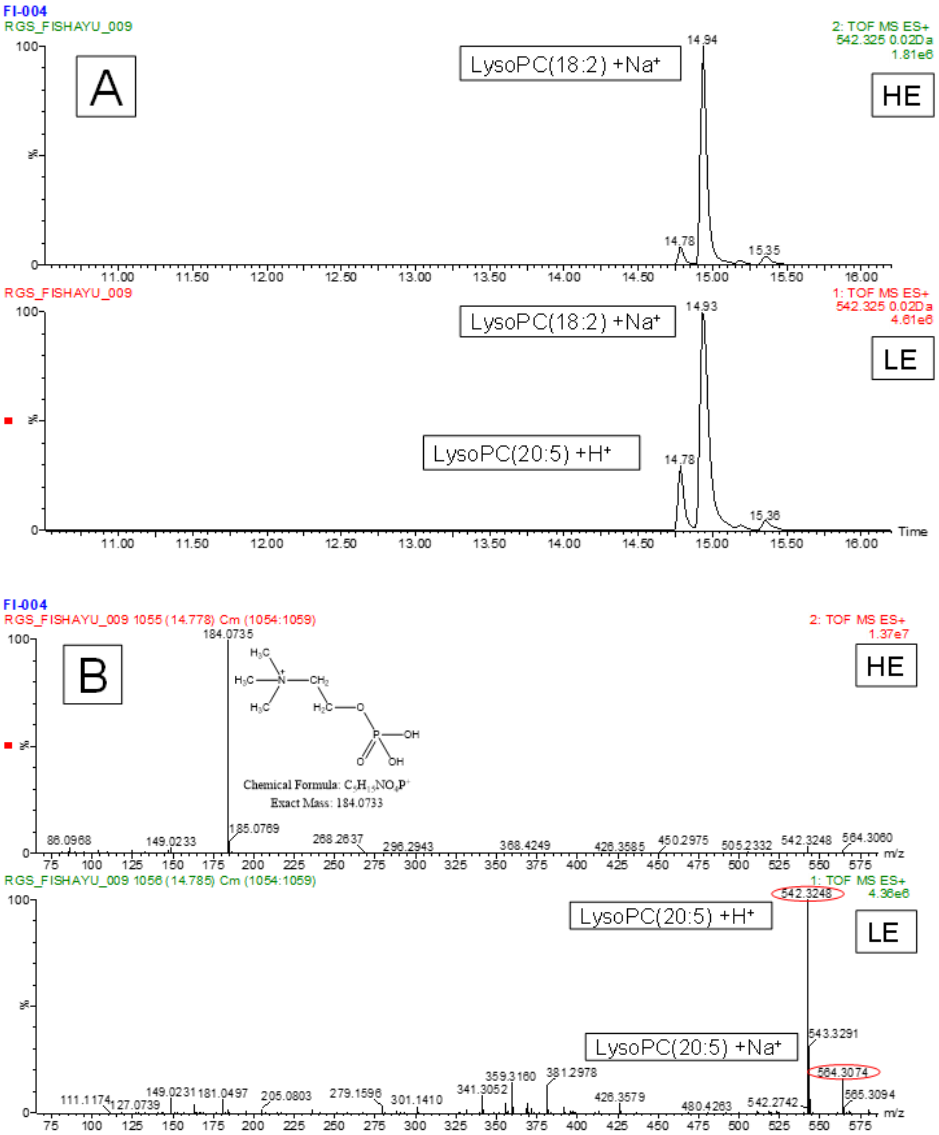


Fig. S4: Chromatogram (A) and mass spectra (B) for LysoPC(20:5) in both low (bottom) and high (top) energy functions in MS^E acquisition mode. Protonated and sodium-adduct molecule appears in the LE function (B at bottom) and LysoPC are observed in HE function (B at top)

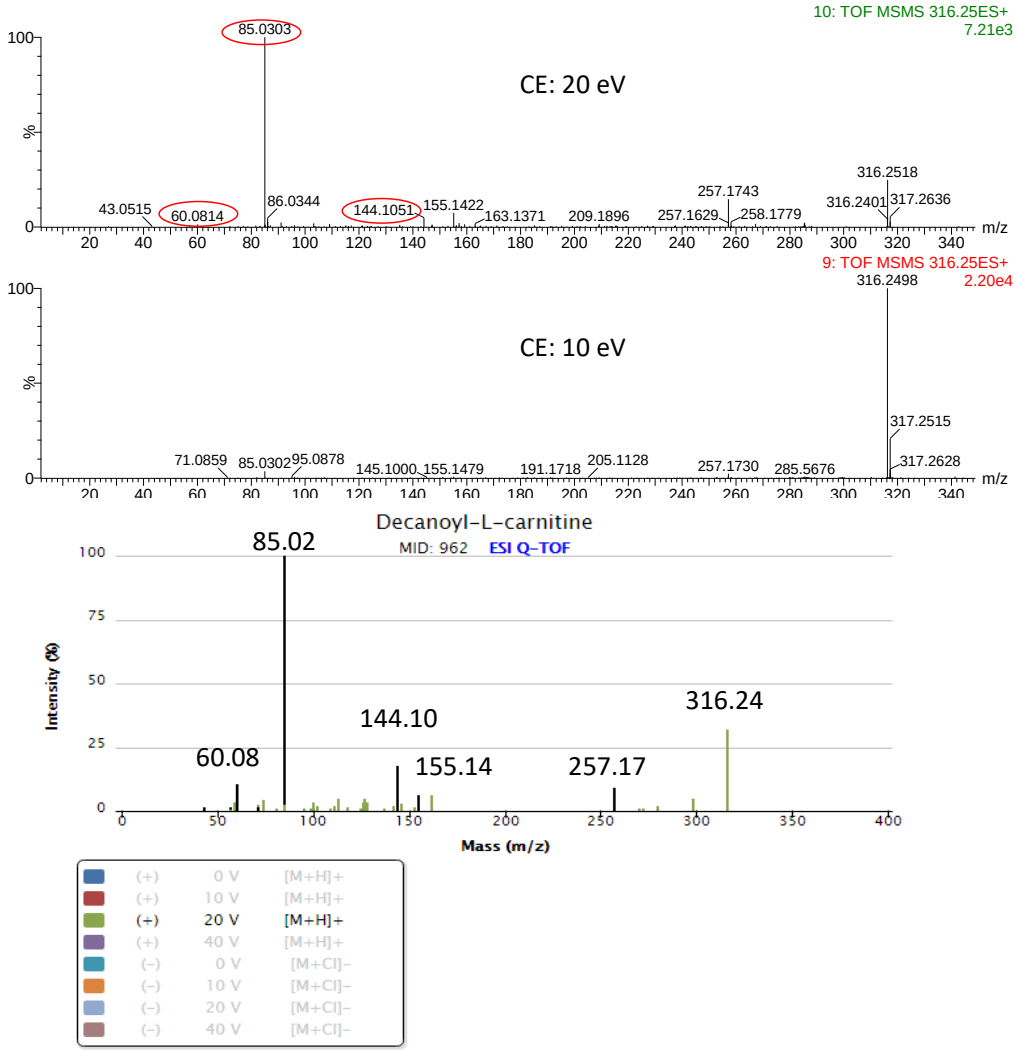


Fig. S5: A: MS/MS spectra at 10 eV and 20 eV collision energy of compound 2 (in table 3). Ions marked in red are specific for carnitine related compounds. B: METLIN spectra for Decanoyl-L-carnitine

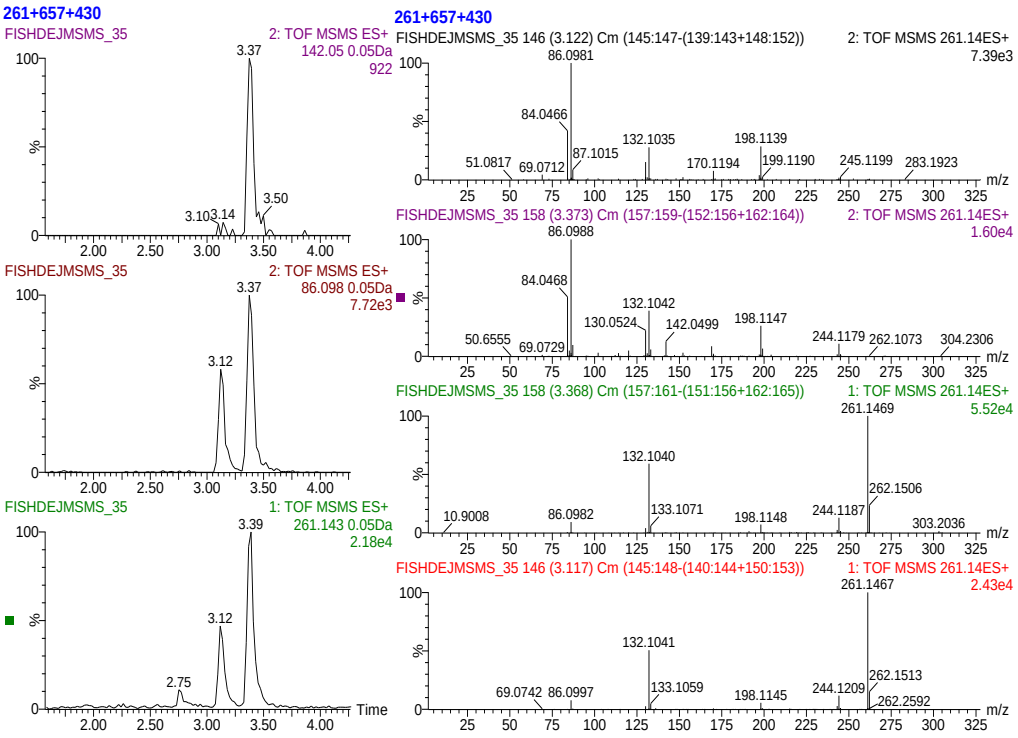


Fig. S6: γ-Glu-Ile and γ-Glu-Leu fragmentation patterns which were employed to assign both isomers. A1: XIC at m/z 261.143 for a MS^E injection, A2: XIC from a MS/MS experiment at 20 eV for a common fragment ion of both compounds (γ-Glu-Ile and γ-Glu-Leu). A3: XIC from a MS/MS experiment at 20 eV for a fragment ion only occurred in γ -Glu-leu. B: MS/MS spectrum isolating m/z 261.1 at 10 eV for γ-Glu-Ile (B1), γ-Glu-Leu (B2) and at 20 eV for γ-Glu-Leu (B3), γ-Glu-Ile (B4) which shows the specific fragment ions (B3). C: METLIN MS/MS spectrum for γ-Glu-Leu. D: METLIN MS/MS spectrum for γ-Glu-Ile.

II.2: Artículo científico 2

Aquaculture 498 (2019) 503–512

Contents lists available at ScienceDirect

Aquaculture

journal homepage: www.elsevier.com/locate/aquaculture

Contributions of MS metabolomics to gilthead sea bream (*Sparus aurata*) nutrition. Serum fingerprinting of fish fed low fish meal and fish oil diets

Rubén Gil-Solsona[a], Josep Alvar Calduch-Giner[b], Jaime Nácher-Mestre[a,b,c], Leticia Lacalle-Bergeron[a], Juan Vicente Sancho[a], Félix Hernández[a], Jaume Pérez-Sánchez[b,*]

[a] Research Institute for Pesticides and Water (IUPA), Avda. Sos Baynat, s/n. University Jaume I, 12071 Castellón, Spain.
[b] Institute of Aquaculture Torre de la Sal (IATS, CSIC), 12595 Ribera de Cabanes, Castellón, Spain.
[c] University Center EDEM, Muelle de la Aduana s/n, 46024, Valencia. Spain

ARTICLE INFO

Keywords:
Fish nutrition
Liquid chromatography
Mass spectrometry
Metabolomics
Vitamins
Microbiota
Plant-based diets

ABSTRACT

The aim of this study was to evaluate the impact of fish meal (FM) and fish oil (FO) replacement by plant proteins and oils in the serum metabolome of two-year old gilthead sea bream (*Sparus aurata*) fed from early life stages with control and experimental diets. Randomly selected fish were overnight sampled and clotted serum was used for metabolomics fingerprinting by means of ultra-high performance liquid chromatography coupled to quadrupole time-of-flight mass spectrometry. > 12,500 different *m/z* ions were detected, and Partial Least Squares-Discriminant analysis separated fish fed control and plant-based diets, with a 71% of variance explained and 44% of variance predicted by the two first components. After variable importance in projection (VIP) and Benjamini-Hochberg test correction filtering, 50 endogenous compounds were elucidated as highly discriminant features of dietary treatment. Most of them were lipid-related compounds and reflected the different fatty acid composition of dietary oils, whereas changes in N-acyl taurines, cytidine and nucleoside related compounds would indicate changes in tissue repair and DNA degradation processes. Untargeted analysis also identified some exogenous compounds as markers of marine and vegetable raw materials. In the case of hercynine (antioxidant fungi and mycobacteria product), this was exemplified by a close lineal association between circulating and feed levels. Targeted approaches were focused on vitamins and a clear reduction of B_{12}, indirectly assessed via methylmalonic acid levels, was found in fish fed vegetable diets. Conversely, serum riboflavin (B_2) and pantothenic acid (B_5) levels were consistently increased, which highlighted the close link between nutrition and gut microbiota.

1. Introduction

Current stagnation of fish meal (FM) and fish oil (FO) production from wild fisheries limits further growth of aquaculture (Tacon and Metian, 2015). The most obvious alternatives are the plant ingredients, which are a common practice in salmonids and marine fish to reduce the reliance of European aquaculture on marine fishery resources. Major progress in this way has been achieved within AQUAMAX, and ARRAINA EU projects and data on key performance indicators clearly indicate that alternative feeds with < 7% marine ingredients support the maximum growth of gilthead sea bream (*Sparus aurata*) from early life stages to completion of sexual maturation (Benedito-Palos et al., 2016; Simó-Mirabet et al., 2018). It is also noteworthy that plant-based diets did not have a negative impact on the shelf life of gilthead bream, trout or carp as high quality foods (Grigorakis et al., 2018). Also, both in salmon and gilthead sea bream, no transfer from feeds to edible fillets was found for regulated mycotoxins, pesticides and persistent organic pollutants with the current plant-based diet formulations (Berntssen et al., 2005, 2010; Nácher-Mestre et al., 2009, 2015; Bell et al., 2012; Portolés et al., 2017). However, regardless of fish fatty acid (FA) biosynthetic capabilities, the use of plant-based diets is associated with a reduced content of n-3 long-chain poly-unsaturated FAs (PUFA) in the meat of farmed fish (Benedito-Palos et al., 2009; Liland et al., 2013, Ballester-Lozano et al., 2016; Turchini et al., 2018).

Other drawback effects of plant-based diets in marine farmed fish are related to changes in fish health and stress resilience (Montero and Izquierdo, 2010). Certainly, the magnitude and persistence of high plasma cortisol levels after crowding exposure is increased in juveniles of gilthead sea bream fed vegetable oils (Ganga et al., 2011), although a lower risk of oxidative stress in these challenged fish is also inferred

* Corresponding author.
E-mail address: jaime.perez.sanchez@csic.es (J. Pérez-Sánchez).

https://doi.org/10.1016/j.aquaculture.2018.08.080
Received 22 May 2018; Received in revised form 30 August 2018; Accepted 31 August 2018
Available online 02 September 2018
0044-8486/

Contributions of UHPLC-QTOF MS metabolomics to gilthead sea bream (Sparus aurata) nutrition: Serum fingerprinting of fish fed low fish meal and fish oil diets

Rubén Gil-Solsona[1], Josep Alvar Calduch-Giner[2] ,Jaime Nácher-Mestre[1,2,3], Leticia Lacalle-Bergeron[1], Juan Vicente Sancho[1],Félix Hernández[1], Jaume Pérez-Sánchez[2]

[1] Research Institute for Pesticides and Water (IUPA), University Jaume I, Castellón, Spain.

[2] Institute of Aquaculture Torre de la Sal (IATS, CSIC), Ribera de Cabanes, Castellón, Spain.

[3] University Center EDEM, Muelle de la Aduana s/n, 46024 Valencia. Spain

Abstract

The aim of this study was to evaluate the impact of fish meal (FM) and fish oil (FO) replacement by plant proteins and oils in the serum metabolome of two-year old gilthead sea bream (*Sparus aurata*) fed from early life stages with control and experimental diets. Randomly selected fish were overnight sampled and clotted serum was used for metabolomics fingerprinting by means of ultra-high performance liquid chromatography coupled to quadrupole time-of-flight mass spectrometry. More than 12,500 different m/z ions were detected, and Partial Least Squares-Discriminant analysis separated fish fed control and plant-based diets, with a 71% of variance explained and 44% of variance predicted by the two first components. After variable importance in projection (VIP) and Benjamini-Hochberg test correction filtering, 50 endogenous compounds were elucidated as highly discriminant features of dietary treatment. Most of them were lipid-related compounds and reflected the different fatty acid composition of dietary oils, whereas changes in N-acyl taurines, cytidine and nucleoside related compounds would indicate changes in tissue repair and DNA degradation processes. Untargeted analysis also identified some exogenous compounds as markers of marine and vegetable raw materials. In the case of hercynine (antioxidant fungi and mycobacteria product), this was exemplified by a close lineal association between circulating and feed levels. Targeted approaches were focused on vitamins and a clear reduction of B12, indirectly assessed via methylmalonic acid levels, was found in fish fed vegetable diets. Conversely, serum

riboflavin (B2) and pantothenic acid (B5) levels were consistently increased, which highlighted the close link between nutrition and gut microbiota.

Introduction

Current stagnation of fish meal (FM) and fish oil (FO) production from wild fisheries limits further growth of aquaculture (*Tacon and Metian, 2015*). The most obvious alternative feeds are the plant ingredients, which are a common practice in salmonids and marine fish to reduce the reliance of European aquaculture on marine fishery resources. Major progress in this way has been achieved within AQUAMAX, and ARRAINA EU projects and data on key performance indicators clearly indicate that alternative feeds with less than 7% marine ingredients support the maximum growth of gilthead sea bream (*Sparus aurata*) from early life stages to completion of sexual maturation (*Benedito-Palos et al., 2016; Simó-Mirabet et al., 2018*). It is also noteworthy that plant-based diets did not have a negative impact on the shelf life of gilthead bream, trout or carp as high quality foods (*Grigorakis et al., 2018*). Also, both in salmon and gilthead sea bream, no transfer from feeds to edible fillets was found for regulated mycotoxins, pesticides and persistent organic pollutants with the current plant-based diet formulations (*Berntssen et al., 2005, 2010; Nácher-Mestre et al., 2009; 2015; Bell et al., 2012; Portolés et al., 2017*). However, regardless of fish fatty acid (FA) biosynthetic capabilities, the use of plant-based diets is associated with a reduced content of n-3 long-chain poly-unsaturated FAs (PUFA) in the meat of farmed fish (*Benedito-Palos et al., 2009; Liland et al., 2013, Ballester-Lozano et al., 2016; Turchini et al., 2018*).

Other drawback effects of plant-based diets in marine farmed fish are related to changes in fish health and stress resilience (*Montero and Izquierdo, 2010*). Certainly, the magnitude and persistence of high plasma cortisol levels after crowding exposure is increased in juveniles of gilthead sea bream fed vegetable oils (*Ganga et al., 2011*), although a lower risk of oxidative stress in these challenged fish is also inferred (*Pérez-Sánchez et al., 2013b*). However, below the threshold level for the theoretical requirements in essential FAs, high inclusion levels of vegetable oils allow a faster disease progression in juveniles of gilthead sea bream challenged with the intestinal parasite Enteromyxum leei (*Estensoro et al., 2011; Calduch-Giner et al., 2012*). A possible cause are the nutritionally-mediated changes on the intestinal profile of mucins, mucosal immunoglobulins (IgT) and other immune-

relevant genes of either diagnostic or predictive value (C*alduch-Giner et al., 2012; Pérez-Sánchez et al., 2013c; Piazzon et al., 2016*), which revealed a pro-inflammatory condition affecting also the integrity of the intestinal barrier (*Estensoro et al., 2016*) and the composition of gut microbiota and intestinal mucus proteome (*Piazzon et al., 2017*). From these studies, however, it was also conclusive that most of these disturbing effects are reversed by the supplementation of plant-based diets with sodium butyrate, resulting in improved diseases outcomes in fish challenged with *E. leei* and the bacteria *Photobacterium damselae subsp. piscicida* (*Piazzon et al., 2017*). Experimental evidence also indicates that diets enriched with medium-chain fatty acid salts (sodium heptanoate, sodium dodecanoate) have a positive impact on feed intake and energy metabolism of juvenile fish reared under sub-optimal conditions (*Simó-Mirabet et al., 2017; Martos-Sitcha et al., 2018*), although possible mechanisms still await full elucidation.

Very often, the application of targeted analyses is the prevailing strategy for qualitative and quantitative detection of different biomarkers. However, this strategy restricts the possibilities to detect other unpredictable effects that could result directly or indirectly from the changes in diet composition. This limitation has encouraged the development and application of new and powerful analytical approaches to face the complexity of this problem and to improve the chance to detect unanticipated effects. Currently a promising new "omic" approach is metabolomics, which aims to use profiles of low-molecular weight metabolic entities (usually < 1,000 Da) to identify biomarkers indicative of specific conditions and particular metabolic pathways. The novelty of this approach in aquaculture research is highlighted in the review article of Alfaro and Young (2018). In particular, nuclear magnetic resonance (NMR)-based lipid fingerprinting allows to precisely classify wild and farmed gilthead sea bream based on their muscle lipid composition (*Melis et al., 2014*). In another gilthead sea bream study, Robles et al. (2013) measured over 80 metabolites from fish intestine samples using a high-performance liquid chromatography-mass spectroscopy (HPLC–MS) platform. Although both analytical platforms rely on wide-untargeted approaches, MS allows retrospective analysis and a higher sensitivity and resolution power (*Castro-Puyana and Herrero, 2013*). Indeed, we have detected more than 15,000 m/z ions in the serum of gilthead sea bream by means of ultra-high performance liquid chromatography (UHPLC) and high resolution MS (HRMS) (*Gil-Solsona et al., 2017*). The same platform has been used in the present study to analyse fish from the eight-months feeding

trial of Benedito-Palos et al. (2016). That study was prolonged, and herein data on wide- and targeted-serum metabolome were used to underline the effects of alternative feeds in two-year old fish fed experimental diets from early life stages.

Materials & methods

Reagents and chemicals

HPLC-grade methanol (MeOH), HPLC-supergradient acetonitrile (ACN), sodium hydroxide (>99%), ammonium hydroxide (NH_4OH) and ammonium acetate (NH_4Ac) were obtained from Scharlab (Barcelona, Spain). HPLC-grade water was obtained from a Milli-Q water purification system (Millipore Ltd., Bedford, MA, USA). Leucine-enkephalin (mass-axis calibration), formic acid (mobile phase modifier), N,N-dimethyl L-histidine (reagent grade), methyl iodine (reagent grade) and tetrabuthylammonium acetate (reagent grade) were purchased from Sigma-Aldrich (Saint Louis, MO, USA).

Diets

Four experimental diets were formulated and produced by BioMar (Brande, Denmark). All diets were isonitrogenous, isolipidic and isoenergetic and met all known nutritional requirements of gilthead sea bream. FM was included at 23% in the D1 (control) diet and at 3% in the other three experimental diets (D2, D3 and D4). Fish hydrolysate (CPSP) was added at 2% in all diets. Added oil was either FO (D1 diet) or a blend of vegetable oils (1:1 ratio of rapeseed oil: palm oil), replacing 58% (D2 diet) and 84% (D3 and D4 diets) FO. A commercial butyrate preparation (BP-70®, NOREL) was added to the D4 diet at 0.4%. All diets contained histidine (0.14%), antioxidants (0.045%) and a mineral-vitamin mix (0.5%). Lysine, methionine, choline, lecithin and monocalcium phosphate were balanced in D2, D3 and D4 diets to the values of the control diet (**Table 1**).

Table 1. Ingredients and chemical composition of experimental diets.

Ingredient (%)	Diet			
	D1	D2	D3	D4
Fish meal	23.0	3.0	3.0	3.0
Fish hydrolysate (CPSP)	2.0	2.0	2.0	2.0
Soya protein	16.0	25.0	25.0	25.0
Corn gluten	15.0	25.0	25.0	25.0
Wheat gluten	4.0	7.3	7.3	7.3
Rapeseed cake	12.0	9.7	9.9	9.9
Wheat	11.08	6.80	6.64	6.24
Fish oil	15.60	6.56	2.50	2.50
Rapeseed oil	0.0	4.4	6.5	6.5
Palm oil	0.0	4.4	6.5	6.5
Monocalcium phosphate	0.303	2.097	2.097	2.097
Histidine	0.136	0.136	0.136	0.136
Mineral Vitamin mix [a]	0.5	0.5	0.5	0.5
Cholesterol	0.113	0.113	0.113	0.113
Amino-acid and micronutrient mix [b]	0.20	2.92	2.74	2.74
Antioxidants	0.045	0.045	0.045	0.045
Yttrium	0.03	0.03	0.03	0.03
Butyrate (BP-70)	0.0	0.0	0.0	0.4
Proximate composition				
Dry matter (DM, %)	91.65	91.79	91.80	92.34
Crude protein (%DM)	45.48	46.73	46.12	46.03
Crude fat (% DM)	19.80	19.56	20.13	19.40
EPA + DHA (% DM)	2.90	1.38	0.67	0.63

[a] Supplied the following (g/kg mix, except as noted): calcium 689, sodium 108, iron 3, manganese 1, zinc 1, cobalt 2 mg, iodine 2 mg, selenium 20 mg, molybdenum 32 mg, retinyl acetate 1, DL-cholecalciferol 2.6, DL-α tocopheryl acetate 28, menadione sodium bisulphite 2, ascorbic acid 16, thiamin 0.6, riboflavin 1.7, pyridoxine 1.2, vitamin B_{12} 50 mg, nicotinic acid 5, pantothenic acid 3.6, folic acid 0.6, and biotin 50 mg.

[b] Contains methionine, lysine, choline, and lecithin.

Animal care and sampling

Juvenile fish (15 g initial average body weight) of Atlantic origin (Ferme Marine de Douhet, Ile d'Oléron, France) were fed control and experimental diets in the indoor experimental facilities of the Institute of Aquaculture Torre de la Sal (IATS-CSIC, Spain). Fish were allocated in 2,500 L tanks in triplicated groups (150 fish/tank), and each one was fed one of the experimental diets for 16 months from May 2013 to August 2014. The number of fish per tank was progressively reduced by periodical samplings, maintaining the rearing density below 15 kg/m3. Oxygen content of outlet water remained higher than 75% saturation and day-length and water temperature followed natural changes at IATS-CSIC latitude (40° 5'N; 0° 10'E). At time of sampling, actively fed fish (10 fish per diet) were sampled following overnight fasting for blood and tissue collection. Liver and visceral adipose tissue were extracted and weighed. Blood was taken from caudal vessels with vacutainer tubes with a clot activator, allowed to clot for 30 min at room temperature, and then centrifuged at 1,300 g for 10 min. The obtained samples were stored at -20ºC until analysis.

All procedures were approved by the IATS Ethics and Animal Welfare Committee according to national (*Royal Decree RD53/2013*) and EU legislation (*2010/63/EU*) on the handling of animals for experiments.

HPLC-HRMS

The analytical procedure was similar to that described elsewhere by Gil-Solsona et al. (2017). Briefly, serum samples were deproteinized with ACN and one supernatant aliquot was used for hydrophilic interaction liquid chromatography (HILIC). Another aliquot was evaporated to dryness and re-dissolved in MeOH 10% for reversed phase (RP) analysis. Quality control (QC) samples were prepared by pooling 50 µL of each sample extract. Extracts (10 µL) were injected in HILIC and RP in both positive and negative ionization modes (0.7 kV and 1.5 kV capillary voltages, respectively) in a hybrid quadrupole time-of-flight mass spectrometer (Xevo G2 QTOF, Waters, Manchester, UK) with a cone voltage of 25 V, using nitrogen as both desolvation and nebulizing gas.

The HILIC separation was performed using a mix of ACN:H_2O (95:5, v/v) as weak mobile phase (A) and H_2O as strong mobile phase (B) both in 0.01% formic acid (HCOOH) and 10 mM NH_4Ac. The percentage of B was changed as follows: 0 min, 2%; 1.5 min, 2%; 2.5 min, 15%; 6 min, 50%; 7.5 min, 75%; and finally at 7.51 min, 2%, with a total run time of 10 min, for both ESI+ and ESI-. For RP separation, the weak mobile phase (A) was H_2O with 0.01% HCOOH and the strong mobile phase (B) was MeOH with 0.01% HCOOH. The B percentage was changed from 10% at 0 min, to 90% at 14 min, 90% at 16 min and 10% at 16.01 min, with a total run time of 18 min for both ESI+ and ESI-. In order to obtain a better resolution among isomers of free FAs and phospholipids, aliquots of RP samples were fortified at 50 mM with tetrabuthylammonium acetate (TBA) and injected with the following gradient: A: H_2O 0.01% HCOOH, B: MeOH 0.01% HCOOH; The percentage of B was maintained at 70% during the first 5 min and changed from 70% at 5 min, to 80% at 8 min, 85% at 12 min, 95% at 15 min, 100% at 22 min and 70% again at 22.01 min with a total run time of 24 min for both ESI+ and ESI-.

Untargeted Data Processing

LC-MS data were processed using XCMS R package (*https://xcmsonline.scripps.edu/*) with Centwave algorithm for peak picking (peak width from 5 to 20 s, S/N ratio higher than 10 and mass tolerance of 15 ppm), followed by retention time alignment, peak area normalization (mean centering), log 2 applying (to avoid heteroscedasticity) and Pareto scaling. For elucidation purposes, fragmentation spectra of features of interest were compared with reference spectra databases (*METLIN, http://metlin.scripps.edu; Human Metabolome DataBase, http://www.hmbd.ca; MassBank, http://www.massbank.eu*). For unassigned metabolites, in silico fragmentation software (MetFrag, *http://msbi.ipb-halle.de/MetFrag*), with subsequent searches through Chemspider (*http://www.chemspider.com*) and PubChem (*https://pubchem.ncbi.nlm.nih.gov*) chemical databases, was employed.

Targeted analysis

The retrospective analysis of data acquired in MSE mode served for the refined search of additional relevant compounds. This procedure consisted in the search of the m/z ratio (parent ions) of the metabolites of interest in the LE function, as well as product ions obtained from MS/MS spectrum online databases (*METLIN and Human Metabolome DataBase*) in the HE function. In the case of vitamins and related-compounds, fat-soluble vitamins were not directly analysable in serum, and their related metabolites were analysed as retinol phosphate for vitamin A, 25-hydroxyvitamin D3 for vitamin D3, α-Carboxyethylhydroxychroman for vitamin E and menaquinone for vitamin K2 (*Tai et al., 2010; Lebold et al., 2012; Karl et al., 2014*). Water-soluble vitamins were directly analysed (B1, B2, B5, B6, B7 and C) with the exception of B12, which was indirectly assayed as methylmalonic acid (MMA) (*Lewerin et al., 2003*).

Targeted analysis was also applied for hercynine, a betaine compound synthetized by fungi and mycobacteria. This exogenous compound was analysed in feeds and serum samples, using a hercynine standard synthetized as described elsewhere (*Khonde and Jardine, 2015*). In the case of feed samples, the analytical protocol included a polar extraction procedure previously employed in our laboratory for animal by-products (*Nácher-Mestre et al., 2016*). Briefly, 2.5 g of feeds were extracted with 10 mL H_2O:ACN (20:80) 0.1% HCOOH, centrifuged and supernatant (5 mL) was passed through an OASIS WCX SPE cartridge previously cleaned with 6 mL MeOH and 6 mL of Milli-Q H_2O. Sample was loaded, cleaned with 6 mL of MeOH:H_2O (1:1) and finally eluted in 3 mL of 2% formic acid in methanol. The feeds samples were then lead to dryness and diluted in 200 µL of Milli-Q H_2O to continue with MS analysis.

Statistical analysis

Data on growth performance and targeted analysis were analysed by one-way ANOVA followed by the Student Newman–Keuls test ($P < 0.05$). After data preprocessing of untargeted metabolomics, multivariate analysis was performed to find discriminative features among groups by means of the EZ-Info software (Umetrics, Sweden). First, Principal Component Analysis (PCA) was used to ensure the absence of outliers and the correct classification of QCs after normalization. Then, all the four experimental groups were joined in a single file and Partial Least Squares-Discriminant Analysis (PLS-DA) was conducted to maximize the separation of dietary groups. The contribution of m/z features to the PLS-DA model was assessed by means of variable importance in projection (VIP) measurements. A VIP score > 1 was considered an adequate threshold to determine discriminant variables in the PLS-DA model (*Wold et al., 2001; Li et al., 2012; Kieffer et al., 2016*). Additionally, orthogonal PLS-DA (*Wiklund et al., 2008*) with a high threshold (P [corr] > 0.7) was carried out to highlight the most discriminant compounds. To end, differences in normalized peak areas of m/z features were analysed by One-way ANOVA followed by Benjamini-Hochberg multiple testing correction analysis.

Results and Discussion

Fish condition

In the previous study of Benedito-Palos et al. (2016), data on key performance indicators and gene expression of growth-related markers in liver and skeletal muscle highly supported the suitability of FM/FO replacement by plant ingredients. In agreement with this, when fish coming from this initial trial were randomly sampled for serum metabolomics fingerprinting, all fish showed a similar average body weight ranging between 577 and 612 g (**Table 2**).

Table 2. Biometry of sampled gilthead sea bream fed experimental diets. Values are the mean ± SEM (n= 10).

	D1	D2	D3	D4	P-value (ANOVA)
Body weight (g)	611.95 ± 24.2	587.40± 25.8	580.8± 10.7	577.6 ± 21.0	0.679
Liver weight (g)	7.33 ± 0.33	7.42± 0.64	8.06 ± 0.38	7.38 ± 0.38	0.855
Mesenteric fat (g)	13.80 ± 2.18	11.89 ± 2.16	10.61 ± 1.41	10.38 ± 1.50	0.546
HSI (%)[1]	1.20 ± 0.05	1.27 ± 0.06	1.39 ± 0.06	1.28 ± 0.06	0.124
MSI (%)[2]	2.20 ± 0.31	2.19 ± 0.28	1.80 ± 0.20	1.79 ± 0.25	0.673

[1]Hepatosomatic index = (100 x liver weight) / fish weight.

[2]Mesenteric fat index = (100 x mesenteric fat) / fish weight.

Likewise, hepatosomatic index (HSI) and mesenteric fat index (MSI) remained mostly within the normal range of variation for the class of fish size and season. This revealed a lack of impact of dietary treatment upon body fat storage or tissue lipid trafficking, which are now recognized as clear signs of essential FA deficiencies in gilthead sea bream (*Pérez-Sánchez et. al., 2013a; Ballester-Lozano et al., 2015*). Despite this, integrative omics approaches combining transcriptomics, proteomics and microbiome analyses highlighted a pro-inflammatory phenotype, with changes in the integrity of the epithelial intestinal barrier and diseases outcomes when fish fed plant-based diets are challenged with bacteria and enteric parasites (*Estensoro et al., 2016; Piazzon et al., 2016; 2017*). Recently, it has also been proven that plasma levels of sex steroids and the male-female sex reversal through the life cycle of gilthead sea bream are differentially regulated in fish fed marine and vegetable diets (*Simo-Mirabet et al., 2018*), Nevertheless, sex steroids (testosterone, 11-ketotestosterone, 17ß-estradiol) cannot be considered a major discriminating factor in this study, since their plasmatic concentrations increase gradually through gametogenesis in concomitance with gonadal growth, decreasing abruptly thereafter. Accordingly, circulating sex steroids were almost undetectable in our

experimental setup using fish sampled out of the reproductive period, which normally extents for gilthead sea bream from October to March in our latitude (*Chaoui et al., 2006; Hadj-Taieb et al., 2013*). In any case, our methodology allowed a wide-screening approach, and a total of 12,982 m/z features (ions) were obtained in all four acquisition modes (RP and HILIC in both ionization modes ESI+ and ESI-). These numbers are comparable to those previously reported for fed and fasted juveniles of gilthead sea bream, using the same UHPLC-HRMS platform (*Gil-Solsona et al., 2017*). Of course, not all features corresponded to a single compound, but the number of detectable ions (13,000-15,000) was high enough to have a wide-representation of the serum fish metabolome. Indeed, the number of different compounds in animal biofluids is estimated to be more than 8,000 (*Kałużna-Czaplińska et al., 2014*), with around 4,500 in human blood according to the Human Metabolome DataBase (*Wishart et al., 2013*).

Untargeted fingerprinting: multivariate analysis

One-way ANOVA was suitable to detect a wide-range of changes in circulating metabolites with more than 5,000 different ions when comparing control and extreme D3/D4 groups ($P < 0.05$), but these numbers were drastically reduced after filtering with Benjamini-Hochberg for false positive corrections (**Fig. 1A**). Thus, the number of ions with a different abundance ranged between 451 and 2,929 when comparisons were made between D1 and D2 fish; and D1 and D4 fish, respectively. However, only four individual features were different between groups D3 and D4, which was indicative that the source of variation due to butyrate supplementation was very low in comparison to that of the replacement of FM and FO with plant ingredients.

A

Group name	D2	D3	D4
D1	*2245 (451)*	*5203 (2487)*	*5166 (2610)*
D2	---	*4833 (1509)*	*5724 (2929)*
D3	---	---	*1192 (4)*

B

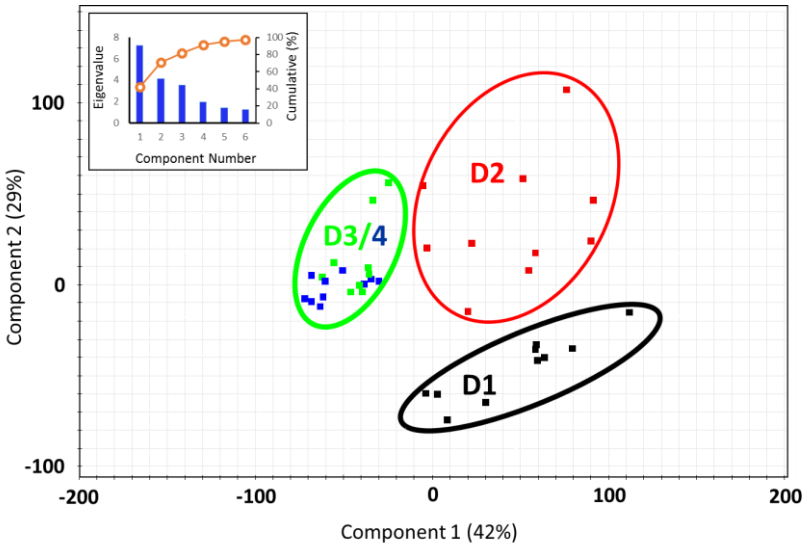


Fig. 1. PLS-DA score plot of acquired data of D1 group individuals (black), D2 (red) and D3/4 (green for D3, blue for D4). Insert is a screen plot of the principal component analysis, showing eigenvalues (blue bars) and cumulative variability explained (orange points) against the number of the principal component.

This was also evidenced by multivariate PLS-DA analysis as many individuals of D3 and D4 groups overlapped in the score plot (**Fig. 1B**). This is the reason because data from fish fed D3 and D4 were joined in the same group (D3/4) for subsequent PLS-DA analyses, where the 71% of variance and 44% of variance was explained or predicted, respectively, by the two first components. The maximum individual variability was achieved within D2 group, but importantly all fish of D1 and D3/D4 groups were correctly classified in the discriminant model. Thus, the maximum separation along both components was found for D1 and D3/4 fish that were distributed along the first (X-axis) and second (Y-axis) component, whereas the separation of D2 and D3/4 fish was only evidenced along the first component reflecting the changes in FO inclusion levels (6.5% D2 diet; 2.50% D3/4 diets). In contrast, the distribution along the second component would primarily reflect the reduced feed intake of FM and fish hydrolysates with inclusion levels of 25.0% in D1 diet and 5.0% in D2, D3

and D4 diets. However, it is noteworthy that the number of features with a P[corr]>0.95 by Orthogonal PLS-DA was reduced to 39, whereas up to 850 ions were identified as clearly discriminant ions in 10-days fasted fish (Gil-Solsona et al., 2017). Therefore, in comparison to short-term fasting, the magnitude of changes induced by dietary treatment herein was relatively low, which explains that elucidation procedures were not restricted to the most discriminant compounds (see below).

Elucidation of untargeted differential compounds

A total of 55 representative compounds with statistically significant changes in abundance after correction for false positives and a VIP score > 1.3 were elucidated (**Table 3**). Most of them were compounds of lipid nature, such as phosphocholines (PC, 24), lysophosphocholines (LysoPC, 10), free FAs (8) and sphingolipids (2). Other compounds with a different abundance were elucidated as N-acyl-taurines (2), cytidine and cytosine nucleosides (4), cysteinolic acid, tauropine, trimethylamine N-oxide (TMAO), arsenobetaine and hercynine. Accordingly, most of these compounds are related to lipid metabolism and highly reflected the decreased unsaturation index of FAs of vegetable oils. Indeed, FO has an elevated content of n-3 LC-PUFAs, whereas vegetable oils are almost devoid of eicosapentaenoic acid (20:5n-3) and docosahexaenoic acid (22:6n-3), which cannot be synthetized at a high rate in marine fish from the C18 PUFA precursor, α-linolenic acid (18:3n-3) (*Tocher, 2015*). In consequence, previous studies in gilthead sea bream clearly indicate that the inclusion of vegetable oils in fish feeds reduced the content of LC-PUFAs and increased that of C18 PUFAs in liver, adipose tissue and muscle fillets, with a selective incorporation of unsaturated FAs in polar lipids (*Izquierdo et al., 2005; Benedito-Palos et al., 2010; 2013*) to preserve and maintain the function of cell membrane surfaces. Especially for fat fish species, most of these changes in the flesh FA composition are highly predictable by means of a dummy regression model (*Ballester-Lozano et al., 2014; 2016*). Less is known about the effects of diet composition on the FA composition of circulating lipids, although overall they also reflect the changes in diet composition (*Laidlaw and Holub, 2003; Lemaitre et al., 2003*) as it was herein the case of circulating PCs, lysoPCs and free FAs. Moreover, the number and degree of these changes in comparison to control group D1 increase with the level of replacement in a dose-dependent manner.

Table 3. Highlighted compounds obtained from untargeted metabolomics. Values are the mean ± SEM (n= 8-10).

	Compound name	Biological process[+]	Chromatography/ ionization mode	Formula	De/protonated molecule *m/z* (mDa)	RT (min)	D2, % CTRL	D3/4, %CTRL	Corrected P-value[‡]	VIP[††]
1	PC(22:6/16:0)	1	RP(spec) / +	$C_{46}H_{80}NPO_8$	806.5701 (+0.1)	18.86	68 ± 5^b	39 ± 7^c	1.63E^{-06}	2.12
2	PC(22:6/18:0)	1	RP(spec) / +	$C_{48}H_{84}NPO_8$	834.6010 (-0.3)	19.86	71 ± 13^b	49 ± 6^c	3.57E^{-06}	1.48
3	PC(22:6/18:3)	1	RP(spec) / +	$C_{48}H_{78}NPO_8$	828.5544 (+0.1)	17.65	94 ± 15^a	49 ± 4^b	1.54E^{-03}	1.47
4	PC(22:6/20:4)	1	RP(spec) / +	$C_{50}H_{80}NPO_8$	854.5700 (+0.1)	19.05	38 ± 3^b	16 ± 2^c	7.16E^{-10}	1.90
5	PC(22:6/20:5)	1	RP(spec) / +	$C_{50}H_{78}NPO_8$	852.5541 (-0.2)	17.54	58 ± 4^b	15 ± 2^c	1.91E^{-09}	1.33
6	PC(20:5/14:0)	1	RP(spec) / +	$C_{42}H_{72}NPO_8$	750.5079 (+0.5)	23.44	32 ± 4^b	13 ± 1^c	1.62E^{-15}	2.03
7	PC(20:5/16:0)	1	RP(spec) / +	$C_{44}H_{78}NPO_8$	780.5532 (-1.1)	18.45	82 ± 8^b	30 ± 5^c	4.31E^{-11}	1.93
8	PC(20:5/16:1)	1	RP(spec) / +	$C_{44}H_{76}NPO_8$	778.5385 (-0.2)	17.88	33 ± 3^b	15 ± 1^c	1.82E^{-15}	1.86
9	PC(20:5/18:0)	1	RP(spec) / +	$C_{46}H_{82}NPO_8$	808.5855 (-0.1)	19.5	64 ± 12^b	39 ± 1^c	8.38E^{-07}	1.96
10	PC(20:5/18:1)	1	RP(spec) / +	$C_{46}H_{80}NPO_8$	806.5700 (0.0)	18.56	107 ± 14^a	80 ± 6^b	6.59E^{-02}	1.47
11	PC(20:5/18:2)	1	RP(spec) / +	$C_{46}H_{78}NPO_8$	804.5541(-0.2)	17.88	51 ± 8^b	20 ± 2^c	4.98E^{-07}	1.72
12	PC(20:5/18:3)	1	RP(spec) / +	$C_{46}H_{76}NPO_8$	802.5277 (-1.0)	18.43	125 ± 18^b	63 ± 12^c	3.84E^{-02}	1.38
13	PC(20:5/20:4)	1	RP(spec) / +	$C_{48}H_{76}NPO_8$	826.5381 (-0.6)	17.25	92 ± 8^a	46 ± 6^b	4.98E^{-04}	1.33
14	PC(20:5/20:5)	1	RP(spec) / +	$C_{48}H_{74}NPO_8$	824.5220 (-1.0)	17.28	43 ± 4^b	9 ± 1^c	3.26E^{-08}	1.66
15	PC(18:2/16:0)	1	RP(spec) / +	$C_{42}H_{80}NPO_8$	758.5701 (+0.1)	19.23	410 ± 33^b	571 ± 74^c	3.07E^{-14}	2.13
16	PC(18:2/18:0)	1	RP(spec) / +	$C_{44}H_{84}NPO_8$	786.6008 (-0.5)	20.34	625 ± 94^b	1689 ± 270^c	1.31E^{-05}	2.12
17	PC(18:2/18:2)	1	RP(spec) / +	$C_{44}H_{80}NPO_8$	782.5711 (+1.1)	18.6	568 ± 68^b	1693 ± 271^c	3.16E^{-08}	1.67
18	PC(18:1/16:0)	1	RP(spec) / +	$C_{42}H_{82}NPO_8$	760.5859 (+0.3)	20.03	151 ± 20^b	193 ± 14^c	2.44E^{-03}	1.31
19	PC(18:1/18:0)	1	RP(spec) / +	$C_{44}H_{86}NPO_8$	788.6191 (+2.2)	21.22	204 ± 33^b	316 ± 35^c	6.10E^{-03}	1.89
20	PC(18:1/18:1)	1	RP(spec) / +	$C_{44}H_{84}NPO_8$	786.6012 (-0.1)	20.31	483 ± 68^b	761 ± 68^c	1.10E^{-06}	2.12
21	PC(18:1/18:2)	1	RP(spec) / +	$C_{44}H_{82}NPO_8$	784.5858 (+0.2)	19.37	766 ± 130^b	1581 ± 285^c	6.29E^{-09}	2.12
22	PC(18:1/18:3)	1	RP(spec) / +	$C_{44}H_{80}NPO_8$	782.5712 (+1.2)	18.6	488 ± 93^b	1419 ± 199^c	6.43E^{-08}	1.67
23	PC(16:0/18:0)	1	RP(spec) / +	$C_{42}H_{84}NPO_8$	762.6013 (0.0)	20.03	517 ± 36^b	1080 ± 119^c	1.74E^{-18}	2.16
24	PC(16:0/18:3)	1	RP(spec) / +	$C_{42}H_{78}NPO_8$	756.5545 (+0.2)	18.5	532 ± 85^b	1317 ± 105^c	3.84E^{-08}	1.96
25	LysoPC(22:6)	1	RP(spec) / +	$C_{30}H_{50}NPO_7$	568.3405 (+0.2)	9.81	77 ± 15^b	60 ± 8^b	1.21E^{-05}	1.55
26	LysoPC(22:5)	1	RP(spec) / +	$C_{30}H_{52}NPO_7$	570.3566 (+0.6)	9.11	143 ± 20^b	157 ± 25^b	7.17E^{-03}	1.37

27	LysoPC(20:5)	1	RP(spec) / +	$C_{28}H_{48}NPO_7$	542.3242 (+0.5)	8.58	88 ± 9[b]	87 ± 7[b]	$4.21E^{-02}$	1.59
28	LysoPC(20:4)	1	RP(spec) / +	$C_{28}H_{50}NPO_7$	544.3386 (-1.7)	8.12	50 ± 10[b]	32 ± 4[b]	$2.93E^{-08}$	133
29	LysoPC(20:2)	1	RP(spec) / +	$C_{28}H_{54}NPO_7$	548.3714 (-0.2)	7.88	294 ± 50[b]	468 ± 47[c]	$1.62E^{-15}$	1.61
30	LysoPC(18:3)	1	RP(spec) / +	$C_{26}H_{48}NPO_7$	518.3248 (+0.1)	10.82	212 ± 17[b]	328 ± 66[c]	$1.69E^{-14}$	1.54
31	LysoPC(18:2)	1	RP(spec) / +	$C_{26}H_{50}NPO_7$	520.3403 (0.0)	9.00	237 ± 28[b]	398 ± 40[c]	$5.03E^{-11}$	1.71
32	LysoPC(18:1)	1	RP(spec) / +	$C_{26}H_{52}NPO_7$	522.3557 (-0.3)	8.41	138 ± 18[b]	195 ± 29[c]	$8.14E^{-07}$	1.73
33	LysoPC(18:0)	1	RP(spec) / +	$C_{26}H_{54}NPO_7$	524.3704 (-1.2)	7.55	113 ± 10[b]	170 ± 19[c]	$3.78E^{-04}$	1.43
34	LysoPC(16:0)	1	RP(spec) / +	$C_{24}H_{50}NPO_7$	496.3402 (-0.1)	10.82	145 ± 26[b]	323 ± 42[c]	$3.33E^{-05}$	1.38
35	FFA(22:6)	2	RP / -	$C_{22}H_{32}O_2$	327.2316 (-0.8)	15.18	85 ± 10[a]	67 ± 13[b]	$3.25E^{-03}$	1.31
36	FFA(20:5)	2	RP / -	$C_{20}H_{30}O_2$	301.2167 (-0.1)	15.17	96 ± 17[a]	77 ± 12[b]	$4.00E^{-03}$	1.55
37	FFA(20:4)	2	RP / -	$C_{20}H_{32}O_2$	303.2316 (-0.8)	15.86	92 ± 12[a]	78 ± 5[b]	$9.00E^{-03}$	1.77
38	FFA(18:4)	2	RP / -	$C_{18}H_{28}O_2$	275.2004 (-0.7)	14.98	51 ± 7[b]	27 ± 5[c]	$7.51E^{-16}$	1.95
39	FFA(18:2)	2	RP / -	$C_{18}H_{32}O_2$	279.2316 (-0.8)	15.91	202 ± 38[b]	295 ± 21[c]	$6.03E^{-09}$	1.95
40	FFA(18:1)	2	RP / -	$C_{18}H_{34}O_2$	281.2472 (-0.9)	15.66	160 ± 18[b]	308 ± 34[c]	$1.32E^{-04}$	1.45
41	FFA(16:1)	2	RP / -	$C_{16}H_{30}O_2$	253.2161 (-0.7)	16.43	103 ± 19[a]	177 ± 25[c]	$3.00E^{-03}$	1.35
42	FFA(16:0)	2	RP / -	$C_{16}H_{32}O_2$	255.2316 (-0.8)	16.43	111 ± 18[a]	136 ± 23[b]	$1.05E^{-02}$	1.40
43	(9-methyl-d19:3) sphingosine	3	RP / +	$C_{19}H_{37}NO_2$	312.2899 (-0.4)	12.32	23 ± 3[b]	16 ± 2[c]	$1.71E^{-11}$	2.11
44	(D14:2)sphingosine	3	RP / +	$C_{14}H_{27}NO_2$	242.2118 (-0.2)	9.17	878 ± 123[b]	148 ± 27[c]	$1.98E^{-04}$	2.06
45	N-Heptadecenoyl taurine	4	RP / -	$C_{19}H_{37}NSO_4$	374.2355 (-1.0)	15.08	52 ± 10[b]	22 ± 2[c]	$6.04E^{-14}$	1.90
46	N-Palmitoleoyl taurine	4	RP / -	$C_{18}H_{35}NSO_4$	360.2209 (0.0)	14.58	47 ± 4[b]	22 ± 3[c]	$2.87E^{-12}$	1.95
47	Cytidine	5	HI / +	$C_9H_{13}N_3O_5$	244.0941 (+0.8)	4.37	235 ± 28[b]	130 ± 21[c]	$1.34E^{-02}$	1.39
48	Cytosine	5	HI / +	$C_4H_5N_3O$	112.0502 (-0.9)	4.35	200 ± 26[b]	120 ± 8[c]	$1.07E^{-02}$	1.78
49	Deoxycytidine	5	HI / +	$C_9H_{13}N_3O_4$	228.0951 (-3.3)	3.48	653 ± 59[b]	190 ± 27[c]	$5.62E^{-03}$	2.27
50	Methylcytosine	5	HI / +	$C_5H_7N_3O$	126.0645 (-2.2)	4.29	734 ± 103[b]	120 ± 24[c]	$3.87E^{-05}$	2.13
51	Cysteinolic acid	6,10	HI / -	$C_3H_9NSO_4$	154.0169 (-0.5)	4.05	18 ± 3[b]	10 ± 2[b]	$6.51E^{-15}$	2.33
52	Tauropine	7,10	HI / -	$C_5H_{11}NSO_5$	196.0286 (+0.6)	2.28	28 ± 4[b]	35 ± 7[b]	$1.74E^{-10}$	2.32
53	TMAO	7,10	HI / +	C_3H_9NO	76.0760 (-0.2)	5.87	53 ± 10[b]	49 ± 9[b]	$2.69E^{-03}$	1.52
54	Arsenobetaine	8,10	HI / +	$C_5H_{11}AsO_2$	179.0040 (-1.3)	5.78	50 ± 6[b]	52 ± 7[b]	$5.00E^{-05}$	1.97
55	Hercynine	9,10	HI / +	$C_9H_{15}N_3O_2$	198.1235 (-0.8)	5.75	737 ± 74[b]	182 ± 15[c]	$3.00E^{-06}$	2.55

1, Phospholipid metabolism; 2, Fatty acid metabolism; 3, Sphingolipid metabolism; 4, N-acyl amino acid metabolism; 5, Pyrimidine metabolism; 6, Bile acid metabolism/algae amino acid; 7, Anaerobic microbial metabolism; 8, Arsenic metabolism; 9, Fungi metabolism; 10, Exogenous compounds.

[‡]ANOVA followed by Benjamini-Hochberg multiple testing correction. [††]Variable importance in projection measurements in PLS-DA.

Sphingolipids, as well as phospholipids, are essential components of all eukaryotic cell membranes with important roles in a variety of biological processes including cell division and cell-to-cell interactions (*Hannun and Obeid, 2018*). In their simplest forms, sphingosine, phytosphingosine, and dihydrosphingosine serve as the backbones upon which further complexity is achieved. For example, phosphorylation of the C1 hydroxyl group yields the final breakdown products and/or the important signalling molecules sphingosine-1-phosphate, phytosphingosine-1-phosphate and dihydrosphingosine-1-phosphate, respectively (*Gault and Obeid, 2010*). In the present study, two sphingosine-related compounds were altered by dietary treatment and the abundance of (9-methyl-d19:3) sphingosine was markedly reduced (D2, 23% control fish; D3, 16% control) by the replacement of marine resources by plant ingredients. Conversely, (d14:2) sphingosine was markedly increased in D2 fish (878% with respect to D1 group) with intermediate values with the extreme diet formulation in D3/4 fish, which suggests that other factors that the simple inclusion level of plant ingredients have effects on sphingolipid metabolism, but we are still far to underline the physiological significance of this finding.

In recent years, a number of studies have demonstrated the essentiality of dietary taurine for many commercially relevant species, especially marine teleosts. Consequently, the removal of taurine-rich dietary ingredients such as FM can induce deficiencies with a wide range of symptoms, including reduced growth and survival, increased susceptibility to diseases and impaired larval developments as reviewed by Salze and Davis (2015). However, the paradigm that taurine is an essential nutrient is not nearly as clear in freshwater species and it is difficult to draw definitive conclusions, although the list of fish species for which taurine is required is increasing. In any case, taurine is well recognized as an essential nutrient in most carnivorous fish, and early studies in gilthead sea bream indicated that low levels of taurine in the pool of muscle free amino acids is associated with growth impairments in fish fed plant protein-based diets (*Gómez-Requeni et al., 2004*). The amides of long-chain FAs with taurine (N-acyl-taurines) are produced via oxidation of bile acid precursors in peroxisomes, and can function as cell signalling molecules with a wide range of biological activities (*Hunt et al., 2012*). N-acyl-taurines have been recently identified in liver and other rodent tissues, and genetic deletion or pharmacological blockage of the serine amidase FA amide

hydrolase (FAAH) causes profound acceleration on wound healing in mouse skin, and repair associated responses in primary cultures of human keratinocytes and fibroblasts (*Sasso et al., 2016*). In the same study, immunofluorescence images of intact mouse skin show that FAAH co-localizes with cytokeratin 10 and filaggrin, two proteins that are expressed by epidermal supra-basal keratinocytes. In a previous study, we have identified the cytokeratin 8 as a good marker of multiple aquaculture stressors (tank shaking, sounds, moving objects into water, water reverse flow and light flashes) in the skin mucus of gilthead sea bream (*Pérez-Sánchez et al., 2017*). The association of cytokeratines with N-acyl taurines has not been established in fish, but we found herein that the concentration of either N-heptadecenoyl-taurine or N-palmitoleoyl-taurine was progressively and consistently reduced with the combined replacement of FM and FO by plant ingredients. This finding opens new research issues in fish nutrition, which would be targeted to alleviate some of the drawback effects of plant-based diets upon the epithelial mucosae of gilthead sea bream, probably mediated by cell renewal or anti-inflammatory processes, as it has been reported for other bioactive compounds, such as butyrate which helps to restore and preserve the integrity and function in gilthead sea bream fed from early life stages with plant-based diets (*Estensoro et al., 2016; Piazzon et al., 2017*). Moreover, experimental evidence indicates that both butyrate and taurine are able to mitigate through different modes of action the intestinal anomalies of European sea bass fed with highly enriched soybean meal diets (*Rimoldi et al., 2016*).

Cytidine and nucleoside related compounds (cytosine, deoxycytidine, methylcytosine) were also clear discriminant factors in our experimental model, and their concentrations were consistently increased in fish fed plant-based diets. Intriguingly this was more evident in the group of fish fed D2 diet (200-734% control fish) that in the extreme D3/4 group (120-190% control fish). Since these compounds originate from dietary sources, from cellular excretion subsequent to RNA turnover, from cytosolic pools of nucleotides, or from degradation of nuclear DNA phagocytized by macrophages (*Holstege et al., 1984*), it is difficult to understand the physiological significance of these findings, although a major source of variation might be related to some kind of cellular DNA instability. Indeed, the highest difference amount control and experimental groups was reported for deoxycytidine and methylcytosine. Degradation of DNA produces deoxycytidine and chemotherapy sharply raises plasma deoxycytidine levels above pretreatment levels (*Cohen et al., 1997*). At the same

time, methylation of cytosines is an important element of epigenetic regulation, and the increased circulating levels of methylcytosine can indicate not only a higher DNA degradation or instability, but also a hyper-methylation at the whole DNA or at specific gene sites. However, this notion needs to be confirmed by more specific assays, because vegetarian life styles are associated with hypo-methylation states (*Geisel et al., 2005*).

Unlike endogenous compounds, the origin and significance of exogenous compounds with a different abundance was easier to trace, being highly informative of the nature and origin of feed ingredients. Accordingly, the replacement of FM by plant ingredients was associated to a decrease of circulating cysteinolic acid, tauropine, TMAO or arsenobetaine. Cysteinolic acid is a non-protein amino acid similar to taurine, detected in gilthead sea bream and red sea bream (*Pagrus major*) as cholesterol-conjugate precursors in the synthesis of bile salts (*Goto et al., 1996; Une et al., 1991*). This amino acid is not synthesized by fish, but it can be easily incorporated in the food chain as some marine seaweed such as Ulva or Enteromorpha contain large amounts (*Ito, 1963*). Likewise, tauropine is an anaerobic end product found in several marine invertebrate phyla, but widely prevalent in marine molluscs (*Venter et al., 2016*). The same for TMAO, a compound found in animals, plants and fungi, but the concentration of TMAO in marine animals significantly exceeds that of other organisms (*Yancey, 2005*). Likewise, arsenobetaine is the arsenic analogue of the quaternary ammonium compound glycine betaine, and marine animals contain very high levels of this compound, non-toxic for human or animals (*Molin et al., 2015; Stiboller et al., 2015*). Its relative contribution of trophic transfer and biotransformation of arsenic derivatives in the arsenobetaine content in fish is still under debate (*Caumette et al., 2012; Popowich et al., 2016*), although from our results it was evident the direct relation between dietary FM and circulating arsenobetaine levels.

Another exogenous compound with a high discriminant value in our experimental model was hercynine. This is an intermediate compound in the synthesis of ergothioneine, a natural antioxidant that is only synthesized by non-yeast fungi, cyanobacteria and actinobacteria (*Fahey, 2001; Pfeiffer et al., 2011*). Therefore, its detectable presence in the serum of fish is indicative of feeding plant ingredients, although its circulating concentration did not parallel the replacement level, being the circulating concentration (arbitrary units) in D2 fish (737 ± 74%) too much higher than that of D3/4

(182 ± 15%) fish. However, when these values were plotted against the relative concentration of hercynine in the diet, a close linear association was found for this compound (**Fig. 2**).

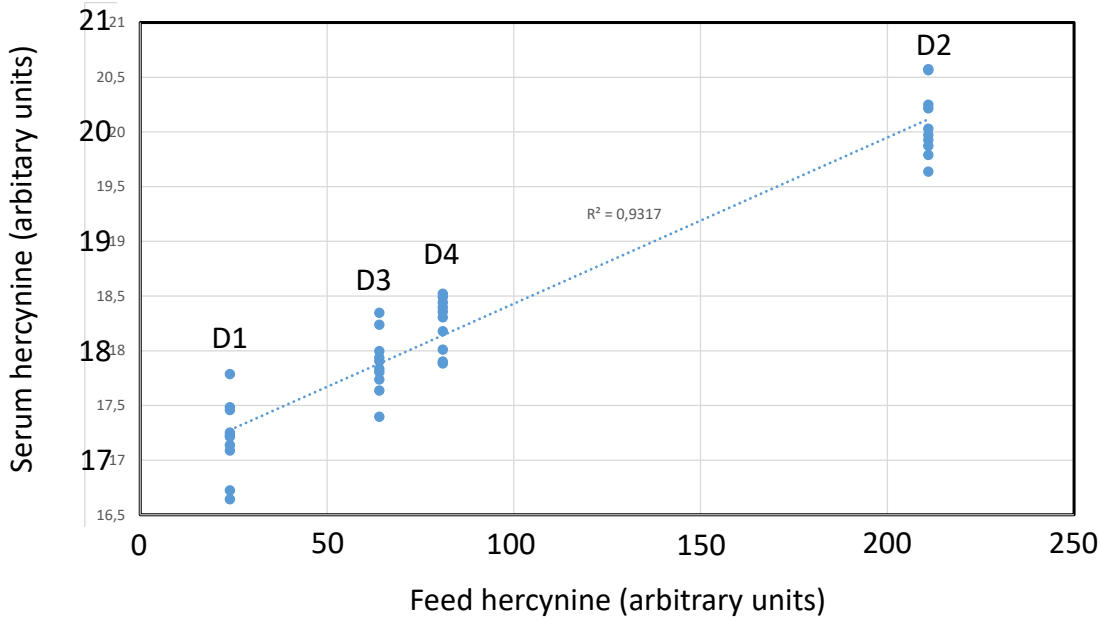


Fig. 2. Correlation plot of hercynine integrated area in feeds (X-axis) and individual serum samples (Y-axis).

Therefore, with the advent of new formulations, hercynine is coming as good biomarker of raw material traceability, but also of proper feed storage and processing of plant-based diets with no fungi/mycobacteria growth.

Targeted vitamin analysis

Vitamins are essential micronutrients that are normally found as precursors of various enzyme reactions in all living cells. However, most of them cannot be synthesized by animals and they need to be obtained exogenously by means of diet fortification, although the use of vitamin-producing microorganisms represents a more natural and consumer-friendly alternative (*Le Blanc et. al., 2013*). In humans, it has been shown that members of the gut microbiota are able to synthesize vitamin K as well as most of the water-soluble B vitamins, such as biotin, cobalamin, folates, nicotinic acid, pantothenic acid, pyridoxine, riboflavin and thiamine (*Hill, 1997*). Unlike dietary vitamins, the predominant uptake of the microbially-produced vitamins occurs in the colon (*Said and Mohammed, 2016*). A similar specialization seems to exist along the digestive tract of fish, as evidenced the microarray gene expression profiling of several genes related to vitamin B12 through the intestine of European sea bass (*Calduch-Giner et al., 2016*). Experimental evidence also indicates that replacement of FM by plant ingredients drives many changes in the micronutrient diet composition, with an important decrease in the content of some vitamins (*NRC, 2011*). In our experimental model, most of the theoretically mineral and vitamin requirements are met in excess by the diet (**Table 1**), but to assess the proper levels of circulating vitamins and vitamin-related compounds, a retrospective (targeted) analysis was conducted by means of the MSE acquisition mode. This approach served to check deficiencies in specific compounds that could have been masked by the astringent Benjamini-Hochberg multiple testing correction in the untargeted approach. Hence, as shown in **Table 4**, the relative concentration of riboflavin (vitamin B2) and pantothenic acid (vitamin B5) were progressively and significantly increased with the replacement of marine sources by plant ingredients in D3/4 fish. Conversely, methylmalonic acid (MMA), used as a biomarker of vitamin B12 deficiency in humans and rodents (*Watanabe et al., 1991; Carmel, 2011*), increased progressively and significantly with the replacement FM/FO by plant ingredients in fish fed D2 and D3/4 diets. The replacement of FM by plant proteins also decreased the concentration of vitamin B12 in muscle and liver tissues of Atlantic cod (*Hansen et al., 2007*), being now well recognized the risk of vitamin B12 deficiency in vegetarian humans (Stabler and Allen, 2004; Allen, 2009). Our targeted approach did not detect additional changes in vitamin condition, although vitamin B7 is markedly reduced by short-term fasting in

gilthead sea bream (*Gil-Solsona et al., 2017*). All this reinforces the importance to define the core microbiota for a given feeding regime and nutritional status, but studies in livestock animal and fish in particular are still in an infancy state to fully understand the complexity of host and gut microbiota interactions.

Table 4. Vitamin and vitamin-related compounds obtained from refined targeted approach. Values are the mean ± SEM (n= 8-10).

Vitamin/vitamin-related compounds		Chormatography/ ionization mode	Formula	De/protonated molecule *m/z*(error mDa)	RT (min)	(%) CTRL D2[†]	(%) CTRL D3/4[†]	P-value (ANOVA)
A	Retinol phosphate	RP/+	$C_{20}H_{31}O_4P$	367.2015 (-2.3)	15.75	140 ± 60[a]	121 ± 63[a]	4.45E-01
B_1	Thiamin	HI/+	$C_{12}H_{16}N_4OS$	265.1118 (-0.5)	5.68	120 ± 18[a]	78 ± 23[a]	2.29E-01
B_2	Riboflavin	RP/-	$C_{17}H_{20}N_4O_6$	375.1299 (-0.6)	4.44	144 ± 67[a]	364 ± 132[b]	1.56E-03
B_5	Pantothenic acid	RP/+	$C_9H_{17}NO_5$	220.1183 (-0.2)	2.04	120 ± 17[a]	146 ± 25[b]	1.98E-02
B_6	Pyridoxine	RP/+	$C_8H_{11}NO_3$	170.0829 (+1.2)	1.72	96 ± 24[a]	104 ± 21[a]	5.96E-01
B_7	Biotin	RP/+	$C_{10}H_{16}N_2O_3S$	245.0955 (-0.5)	5.36	107 ± 19[a]	120 ± 21[a]	3.68E-01
B_{12}	Mehtylmalonic acid (MMA)	RP/-	$C_4H_6O_4$	117.0190 (+0.2)	1.22	195 ± 45[b]	276 ± 35[c]	3.27E-03
C	Dehydroascorbic acid	HI/-	$C_6H_6O_6$	173.0085 (-0.1)	1.12	95 ± 17[a]	122 ± 17[a]	1.03E-01
D_3	25-hydroxyvitamin D_3	RP/+	$C_{27}H_{44}O_2$	401.3412 (-0.8)	13.65	102 ± 39[a]	93 ± 26[a]	3.59E-01
E	α-Carboxyethylhydroxychroman	RP/-	$C_{16}H_{22}O_4$	277.1441 (+0.1)	14.10	106 ± 17[a]	109 ± 15[a]	5.64E-01
K_2	Menaquinone	RP/+	$C_{41}H_{56}O_2$	581.4360 (+0.1)	16.80	72 ± 45[a]	140 ± 54[a]	1.07E-01

[†] Percentage of integrated area for the selected compound as a percentage in fish fed control diet (D1). Compounds with statistical significant differences (P< 0.05) against control fish are in bold.

Conclusions

UHPLC-HRMS approach allowed us to identify a high number of m/z ions in the serum of farmed gilthead sea bream. This was the result of combined targeted and untargeted approaches, which identified a wide-range of endogenous and exogenous compounds with a high discriminant capacity as summarized in **Fig. 3**. Multivariate analyses highlighted a clear separation of fish fed the control and plant-based diets, and the distribution through X-axis and Y-axis evidenced the different effects related to FM or FO replacement by plant proteins and oils. Most of the changes reflected the different FA composition of dietary oils in fish growth at high rates without apparent signs of FA deficiencies. However, N-acyl taurines emerged as target compounds to alleviate some of the negative health effects of plant-based diets. Other metabolite changes (cytidine and nucleoside compounds) highlighted different nutritionally-mediated effects on DNA stability and perhaps methylation levels. Targeted vitamin analysis corroborated the risk of low levels of vitamin B12 in fish fed plant-based diets, whereas other dietary or microbially-produced vitamins were not affected or increased (B2, B5). Lastly, the detection of different exogenous compounds served to trace the use of different raw materials in fish feeds, but also to assess their proper processing and storage.

Author contributions

J.V.S, F.H. and J.P.S conceived and designed the experiments. R.G.S, J.C.G., J.N.M., L.L.B. and J.P.S. performed the experiments. All authors have contributed to analysis of data and the final writing of the paper. All authors have read and approved the final manuscript.

Acknowledgments

The authors acknowledge the financial support of the Spanish MINECO (MI2-Fish, AGL2013- 48560) and Plà de Promoció de la Investigació de la Universitat Jaume I (P1· 1B2015-59). Additional funding was obtained from Generalitat Valenciana, as research groups of excellence (PROMETEO II/2014/023, PROMETEO II/2014/085, Collaborative Research on Environment and

Food-Safety, ISIC/2012/016). This work has been developed in the framework of the Research Unit of Marine Ecotoxicology (IATS-CSIC; IUPA-UJI).

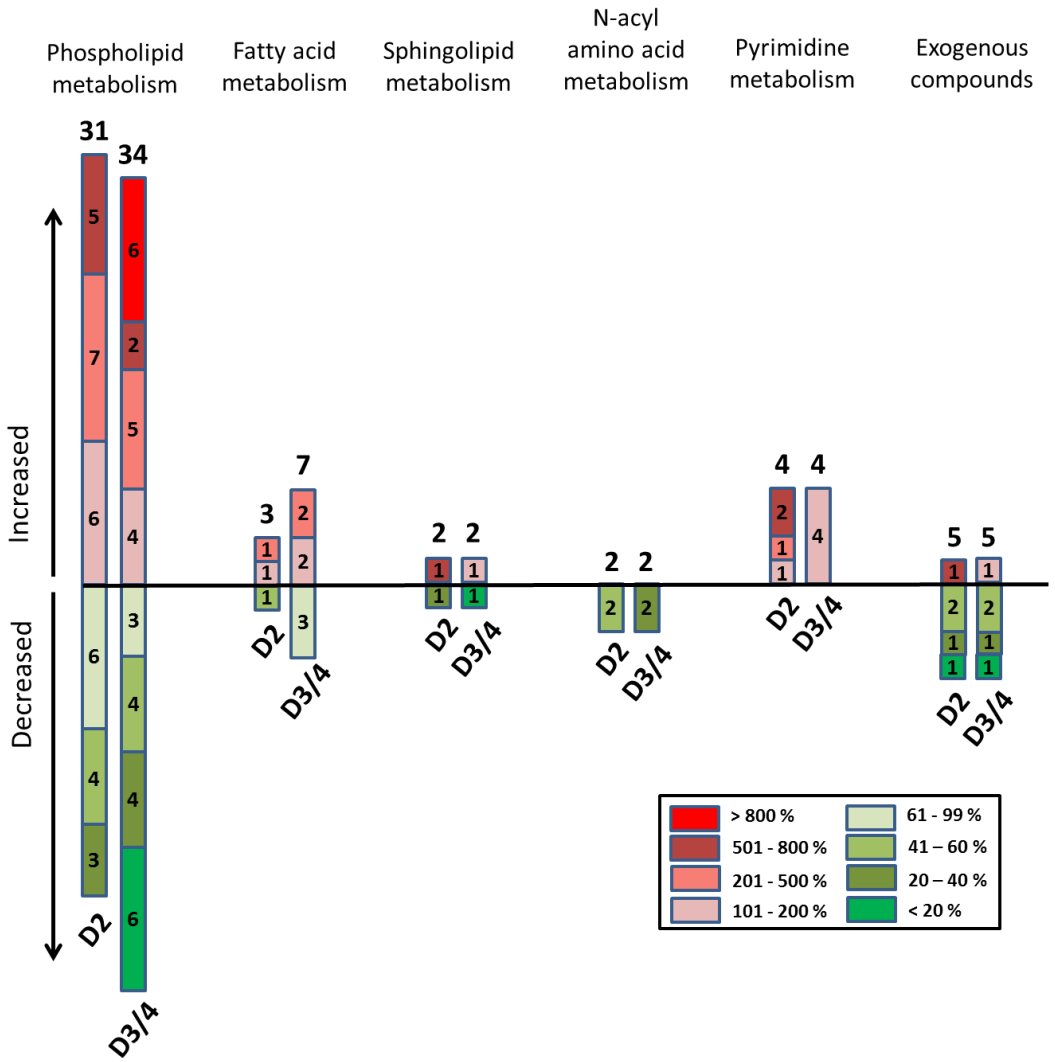


Fig. 3. Integrative profile of differential compounds between D2 and D3/4 compared to control D1 group. Bars show for each dietary group and biological process the number of significantly different (P<0.5, ANOVA followed by Benjamini-Hochberg multiple testing correction) abundant compounds. Colors in each bar indicate the level of change (as % of D1) as indicated in the inbox.

References

Alfaro, A.C., Young, T., 2018. Showcasing metabolomics research in aquaculture: a review. Reviews in Aquaculture 10, 135-152.

Allen, L.H., 2009. How common is vitamin B12 deficiency? Am. J. Clin. Nutr. 89, S693-S696.

Ballester-Lozano, G.F., Benedito-Palos, L., Estensoro, I., Sitjà-Bobadilla, A., Kaushik, S., Pérez-Sánchez J. 2015. Comprehensive biometric, biochemical and histopathological assessment of nutrient deficiencies in gilthead sea bream fed semi-purified diets. British Journal of Nutrition 114, 713-726.

Ballester-Lozano, G.F., Benedito-Palos, L., Mingarro, M., Navarro, J.C., Pérez-Sánchez, J., 2016. Up-scaling validation of a dummy regression approach for predictive modelling the fillet fatty acid composition of cultured European sea bass (Dicentrarchus labrax). Aquaculture Research 47, 1067-1074.

Ballester-Lozano, G.F., Benedito-Palos, L., Riaza, A., Navarro, J.C., Rosel, J., Pérez-Sánchez, J., 2014. Dummy regression analysis for modelling the nutritionally tailored fillet fatty acid composition of turbot and sole using gilthead sea bream as a reference subgroup category. Aquaculture Nutrition 20, 421-430.

Bell, J.G., Dick, J.R., Strachan, F., Guy, D.R., Berntssen, M.H.G., Sprague, M., 2012. Complete replacement of fish oil with a blend of vegetable oils affects dioxin, dioxin-like polychlorinated biphenyls (PCBs) and polybrominated diphenyl ethers (PBDEs) in 3 Atlantic salmon (Salmo salar) families differing in flesh adiposity. Aquaculture 324-325, 118-126.

Benedito-Palos, L., Ballester-Lozano, G.F., Simó, P., Karalazos, V., Ortiz, A., Calduch-Giner, J.A., Pérez-Sánchez, J., 2016. Lasting effects of butyrate and low FM/FO diets on growth performance, blood haematology/biochemistry and molecular growth-related markers in gilthead sea bream (*Sparus aurata*). Aquaculture 454, 8-18.

Benedito-Palos, L., Calduch-Giner, J.A., Ballester-Lozano, G.F., Pérez-Sánchez, J., 2013. Effect of ration size on fillet fatty acid composition, phospholipid allostasis and mRNA expression patterns of lipid regulatory genes in gilthead sea bream (*Sparus aurata*). British Journal of Nutrition 109, 1175-1187.

Benedito-Palos, L., Navarro, J.C., Bermejo-Nogales, A., Saera-Vila, A., Kaushik, S., Pérez-Sánchez, J., 2009. The time course of fish oil wash-out follows a simple dilution model in gilthead sea bream (*Sparus aurata L.*) fed graded levels of vegetable oils. Aquaculture 288, 98-105.

Benedito-Palos, L., Navarro, J.C., Kaushik, S., Pérez-Sánchez, J., 2010. Tissue-specific robustness of fatty acid signatures in cultured gilthead sea bream (*Sparus aurata L.*) fed practical diets

with a combined high replacement of fish meal and fish oil. Journal of Animal Science 88, 1759-1770.

Berntssen, M.H.G., Julshamn, K., Lundebye, A.K., 2010. Chemical contaminants in aquafeeds and Atlantic salmon (*Salmo salar*) following the use of traditional- versus alternative feed ingredients. Chemosphere 78, 637-646.

Berntssen, M.H.G., Lundebye, A.K., Torstensen, B.E., 2005. Reducing the levels of dioxins and dioxin-like PCBs in farmed Atlantic salmon by substitution of fish oil with vegetable oil in the feed. Aquaculture Nutrition 11, 219-231.

Calduch-Giner, J.A., Sitjà-Bobadilla, A., Davey, G.C., Cairns, M.T., Kaushik, S., Pérez-Sánchez, J., 2012. Dietary vegetable oils do not alter the intestine transcriptome of gilthead sea bream (*Sparus aurata*), but modulate the transcriptomic response to infection with Enteromyxum leei. BMC Genomics 470, 13.

Calduch-Giner, J.A., Sitjà-Bobadilla, A., Pérez-Sánchez, J., 2016. Gene expression profiling reveals functional specialization along the intestinal tract of a carnivorous teleostean fish (*Dicentrarchus labrax*). Frontiers in Physiology 7, 359.

Carmel, R., 2011. Biomarkers of cobalamin (vitamin B-12) status in the epidemiologic setting: a critical overview of context, applications, and performance characteristics of cobalamin, methylmalonic acid, and holotranscobalamin II. Am. J. Clin. Nutr. 94, 348S-358S..

Castro-Puyana, M., Herrero, M., 2013. Metabolomics approaches based on mass spectrometry for food safety, quality and traceability, Trends in Analytical Chemistry 52, 74-87.

Caumette, G., Koch, I., Reimer, K.J., 2012. Arsenobetaine formation in plankton: a review of studies at the base of the aquatic food chain. J. Environ. Monitor. 14, 2841-2853.

Chaoui, L., Kara, M.H., Faure, E., Quignard, J.P., 2006. Growth and reproduction of the gilthead seabream *Sparus aurata* in Mellah lagoon (north-eastern Algeria). Scientia Marina 70, 545-552.

Cohen, J.D., Strock, D.J., Teik, J.E., Katz, T.B., Marcel, P.D., 1997. Deoxycytidine in human plasma: Potential for protecting leukemic cells during chemotherapy. Cancer Lett. 116, 167-175.

Estensoro, I., Ballester-Lozano, G.F., Benedito-Palos, L., Grammes, F., Martos-Sitcha, J.A., Mydland, L.-T., Calduch-Giner, J.A., Fuentes, J., Karalazos, V., Ortiz, A., Øverland, M., Sitjà-Bobadilla, A., Pérez-Sánchez, J., 2016. Dietary butyrate helps to restore the intestinal status of a marine teleost (*Sparus aurata*) fed extreme diets low in fish meal and fish oil. PLoS ONE 11, e0166564.

Estensoro, I., Benedito-Palos, L., Palenzuela, O., Kaushik, S., Sitjà-Bobadilla, A., Pérez-Sánchez, J., 2011. The nutritional background of the host alters the disease outcome in a fish-myxosporean system. Veterinary Parasitology 175, 141-150.

Fahey, R.C., 2001. Novel thiols of prokaryotes. Annu. Rev. Microbiol. 51, 333-356.

Ganga, R., Bell, J.G., Montero, D., Atalah, E., Vraskou, Y., Tort, L., Fernández, A., Izquierdo, M.S., 2011. Adrenocorticotrophic hormone-stimulated cortisol release by the head kidney inter-renal tissue from sea bream (*Sparus aurata*) fed with linseed oil and soybean oil. British Journal of Nutrition 105, 238-247.

Gault, C.R., Obeid, L.M., 2011. Still benched on its way to the bedside: sphingosine kinase 1 as an emerging target in cancer chemotherapy. Critical Reviews in Biochemistry and Molecular Biology 46, 342-351.

Geisel, J., Schorr, H., Bodis, M., Isber, S., Hübner, U., Knapp, J.P., Obeid, R., Herrmann, W., 2005. The vegetarian lifestyle and DNA methylation. Clin. Chem. Lab. Med. 43, 1164-1169.

Gil-Solsona, R., Nacher-Mestre, J., Lacalle-Bergeron, L., Sancho, J.V., Calduch-Giner, J.A., Hernández, F., Pérez-Sánchez, J., 2017. Untargeted metabolomics approach for unraveling robust biomarkers of nutritional status in fasted gilthead sea bream (*Sparus aurata*). PeerJ 5, e2920.

Gómez-Requeni, P., Mingarro, M., Calduch-Giner, J., Medale, F., Martín, S.A.M., Houlihan, D.F., Kaushik, S., Pérez-Sánchez, J., 2004. Protein growth performance, amino acid utilisation and somatotropic axis responsiveness to fish meal replacement by plant protein sources in gilthead sea bream (*Sparus aurata*). Aquaculture 232, 493-510.

Goto, T., Ui, T., Une, M., Kuramoto, T., Kihira, K., Hoshita, T., 1996. Bile salt composition and distribution of the D-cysteinolic acid conjugated bile salts in fish. Fish. Sci. 62, 606-609.

Grigorakis, K., Dimitra, K., Corraze, G., Pérez-Sánchez, J., Adorjan, A., Zsuzsanna, J.S., 2018. Impact of diets containing plant raw materials as fish meal and fish oil replacement on rainbow trout (Oncorhynchus mykiss), gilthead sea bream (*Sparus aurata*), and common carp (Cyprinus carpio) freshness. Journal of Food Quality 2018, 1717465.

Hadj-Taieb, A., Ghorbel, M., Hadj-Hamida, N.B., Jarboui O. Sex ratio, reproduction, and growth of the gilthead sea bream, *Sparus aurata* (Pisces: Sparidae), in the Gulf of Gabes, Tunisia. Ciencias Marinas 39, 101-112.

Hannun, Y.A., Obeid, L.M., 2018. Sphingolipids and their metabolism in physiology and disease. Nature Reviews Molecular Cell Biology 19, 175-191.

Hansen, A.C., Rosenlund, G., Karlsen, Ø., Koppe, W., Hemre, G.I., 2007. Total replacement of fish meal with plant proteins in diets for Atlantic cod (Gadus morhua L.). I. Effects on growth and protein retention. Aquaculture 272, 599-611.

Hill, M.J., 1997. Intestinal flora and endogenous vitamin synthesis. Eur. J. Cancer Prev. 6 (Suppl 1), S43-S45.

Holstege, A., Manglitz, D., Gerok, W., 1984. Depletion of blood plasma cytidine due to increased hepatocellular salvage in d-galactosamine-treated rats. Eur. J. Biochem. 141, 339-344.

Hunt, M.C., Siponen, M.I., Alexson, S.E.H., 2012. The emerging role of acyl-CoA thioesterases and acyltransferases in regulating peroxisomal lipid metabolism. Biochimica et Biophysica Acta 1822, 1397-1410.

Ito, K., 1963. Distribution of d-cysteinolic acid in marine algae. Bull. Jap. Soc. Scient. Fish 29, 771-775.

Izquierdo, M.S., Montero, D., Robaina, L., Caballero, R., Rosenlund, G., Ginés, R., 2005. Alterations in fillet fatty acid profile and flesh quality in gilthead seabream (*Sparus aurata*) fed vegetable oils for a long term period. Recovery of fatty acid profiles by fish oil feeding. Aquaculture 250, 431-444.

Kałużna-Czaplińska, J., Jóźwik, J., Żurawicz, E., 2014. Analytical methods used in autism spectrum disorders. Trends in Analytical Chemistry 62, 20-27.

Karl, J.P., Fu, X., Dolnikowski, G.G., Saltzman, E., Booth, S.L., 2014. Quantification of phylloquinone and menaquinones in feces, serum, and food by high-performance liquid chromatography-mass spectrometry. Journal of Chromatography B 963, 128-133.

Khonde, P.L., Jardine, A., 2015. Improved synthesis of the super antioxidant, ergothioneine, and its biosynthetic pathway intermediates. Organic & Biomolecular Chemistry 13, 1415-1419.

Kieffer, D.A., Piccolo, B.D., Vaziri, N.D., Liu, S., Lau, W.L., Khazaeli, M., Nazertehrani, S., Moore, M.E., Marco, M.L., Martin, R.J., Adams, S.H., 2016. Resistant starch alters gut microbiome and metabolomic profiles concurrent with amelioration of chronic kidney disease in rats. Am. J. Physiol. Renal. Physiol. 310, F857-F871.

Laidlaw, M., Holub, B.J., 2003. Effects of supplementation with fish oil-derived n-3 fatty acids and gamma-linolenic acid on circulating plasma lipids and fatty acid profiles in women. Am. J. Clin. Nutr. 77, 37-42.

LeBlanc, J.G., Milani, C., de Giori, G.S., Sesma, F., van Sinderen, D., Ventura, M., 2013. Bacteria as vitamin suppliers to their host: A gut microbiota perspective. Curr. Opin. Biotechnol. 24, 160-168.

Lebold, K.M., Ang, A., Traber, M.G., Arab, L., 2012. Urinary α-carboxyethyl hydroxychroman can be used as a predictor of α-tocopherol adequacy, as demonstrated in the Energetics Study. The American Journal of Clinical Nutrition 96, 801-809.

Lemaitre, R.N., King, I.B., Mozaffarian, D., Kuller, L.H., Tracy, R.P., Siscovick, D.S., 2003. n-3 Polyunsaturated fatty acids, fatal ischemic heart disease, and nonfatal myocardial infarction in older adults: the Cardiovascular Health Study. Am. J. Clin. Nutr. 77, 319-325.

Lewerin, C., Nilsson-Ehle, H., Matousek, M., Lindstedt, G., Steen, B., 2003. Reduction of plasma homocysteine and serum methylmalonate concentrations in apparently healthy elderly subjects after treatment with folic acid, vitamin B12 and vitamin B6: a randomised trial. European Journal of Clinical Nutrition 57, 1426-1436.

Li, H., Ma, M.L., Luo, S., Zhang, R.M., Han, P., Hu, W., 2012. Metabolic responses to ethanol in Saccharomyces cerevisiae using a gas chromatography tandem mass spectrometry-based metabolomics approach. Int. J. Biochem. Cell. Biol. 44, 1087-1096.

Liland, N.S., Rosenlund, G., Berntssen, M.H.G., Brattelid, T., Madsen, L., Torstensen, B.E., 2013. Net production of atlantic salmon (FIFO, fish in fish out < 1) with dietary plant proteins and vegetable oils. Aquaculture Nutrition 19, 289-300.

Martos-Sitcha, J.A., Simó-Mirabet, P., Piazzon, M.C., de las Heras, V., Calduch-Giner, J.A., Puyalto, M., Tinsley, J., Makol, A., Sitjà-Bobadilla, A., Pérez Sánchez, J., 2018. Dietary sodium heptanoate helps to improve feed efficiency, growth hormone status and swimming performance in gilthead sea bream (*Sparus aurata*). Aquaculture Nutrition (in press).

Melis, R., Sanna, R., Braca, A., Bonaglini, E., Cappuccinelli, R., Slawski, H., Roggio, T., Uzzau, S., Anedda, R., 2017. Molecular details on gilthead sea bream (*Sparus aurata*) sensitivity to low water temperatures from 1H NMR metabolomics. Comparative Biochemistry & Physiology Part A Molecular & Integrative Physiology 204, 129-136.

Molin, M., Ulven, S.M., Meltzer, H.M., Alexander, J., 2015. Arsenic in the human food chain, biotransformation and toxicology - review focusing on seafood arsenic. J. Trace Elem. Med. Biol. 31, 249-259.

Montero, D., Izquierdo, M.S., 2010. Welfare and health of fish fed vegetable oils as alternative lipid sources to fish oil, in: Turchini, G., Ng, W., Tocher, D. (Eds), Fish oil replacement and alternative lipid sources in aquaculture feeds. CRC Press, Cambridge, pp. 439-485.

Nácher-Mestre, J., Ibáñez, M., Serrano, R., Boix, C., Bijlsma, L., Lunestad, B.T., Hannisdal, R., Alm, M., Hernández, F., Berntssen, M.H.G., 2016. Investigation of pharmaceuticals in processed animal by-products by liquid chromatography coupled to high-resolution mass spectrometry. Chemosphere 154, 231-239.

Nácher-Mestre, J., Serrano, R., Beltrán, E., Pérez-Sánchez, J., Silva, J., Karalazos, V., Hernández, F., Berntssen, M.H.G., 2015. Occurrence and transfer of mycotoxins in gilthead sea bream and Atlantic salmon by use of novel alternative feed ingredients. Chemosphere 128, 314-320.

Nácher-Mestre, J., Serrano, R., Benedito-Palos, L., Navarro, J.C., López, F.J., Pérez-Sánchez, J., 2009. Effects of fish oil replacement and re-feeding on the bioaccumulation of organochlorine compounds in gilthead sea bream (*Sparus aurata L.*) of market size. Chemosphere 76, 811-817.

NRC, National Research Council, 2011. Nutrient Requirement of Fish and Shellfish, National Academy Press, Washington.

Pérez-Sánchez, J., Benedito-Palos, L., Ballester-Lozano, G.F., 2013a. Dietary lipid sources as a means of changing fatty acid composition in fish: implications for food fortification, in: Preedy, V.R., Srirajaskanthan, R., Patel, V.B. (Eds), Handbook of food fortification and health, from concepts to public health applications, volume 2. Humana Press, New York, pp. 41 – 54.

Pérez-Sánchez, J., Borrel, M., Bermejo-Nogales, A., Benedito-Palos, L., Saera-Vila, A., Calduch-Giner, J.A., Kaushik, S., 2013b. Dietary oils mediate cortisol kinetics and the hepatic expression profile of stress responsive genes in juveniles of gilthead sea bream (*Sparus aurata*) exposed to crowding stress. Comparative Biochemistry and Physiology D 8, 123-130.

Pérez-Sánchez, J., Estensoro, I., Redondo, M.J., Calduch-Giner, J.A., Kaushik, S., Sitjà-Bobadilla, A., 2013c. Mucins as diagnostic and prognostic biomarkers in a fish-parasite model: transcriptional and functional analysis. PLOS One 8, e65457.

Pérez-Sánchez, J., Terova, G., Simó-Mirabet, P., Rimoldi, S., Folkedal, O., Calduch-Giner, J.A., Olsen, R.E., Sitjà-Bobadilla, A., 2017. Skin mucus of gilthead sea bream (*Sparus aurata L.*). Protein mapping and regulation in chronically stressed fish. Frontiers in Physiology 8, 34.

Pfeiffer, C., Bauer, T., Surek, B., Schomig, E., Grundemann, D., 2011. Cyanobacteria produce high levels of ergothioneine. Food Chem. 129, 1766-1769.

Piazzon, M.C., Calduch-Giner, J.A., Fouz, B., Estensoro, I., Simó-Mirabet, P., Puyalto, M., Karalazos, V., Palenzuela, O., Sitjà-Bobadilla, A., Pérez-Sánchez, J., 2017. Under control: how a dietary additive can restore the gut microbiome and proteomic profile, and improve disease resilience in a marine teleostean fish fed vegetable diets. Microbiome 5, 164.

Piazzon, C., Galindo-Villegas, J., Pereiro, P., Estensoro, I., Calduch-Giner, J.A., Gómez-Casado, E., Novoa, B., Mulero, V., Sitjà-Bobadilla, A., Pérez-Sánchez, J., 2016. Differential modulation of IgT and IgM upon parasitic, bacterial, viral and dietary challenges in a perciform fish. Frontiers in Immunology 7, 637.

Popowich, A., Zhang, Q., Le, X.C., 2016. Arsenobetaine: the ongoing mystery. Nat. Sci. Rev. 3, 451-458.

Portolés, T, Ibáñez, M, Garlito, B, Nácher-Mestre, J., Karalazos, V., Silva, J., Serrano, R., Pérez-Sánchez, J., Hernández, F., Berntssen, M.H.G., 2017. Comprehensive strategy for pesticide residue analysis through the production cycle of gilthead sea bream and Atlantic salmon. Chemosphere 179, 242-253.

Raux, E., Schubert, H.L.,Warren, M.J., 2000. Biosynthesis of cobalamin (vitamin B12): A bacterial conundrum. Cell. Mol. Life Sci. 57, 1880-1893.

Rimoldi, S., Finzi, G., Ceccotti, C., Girardello, R., Grimaldi, A., Ascione, C., Terova, G., 2016. Butyrate and taurine exert a mitigating effect on the inflamed distal intestine of European sea bass fed with a high percentage of soybean meal, Fish. Aquat. Sci. 19, 40.

Robles, R., Lozano, A.B., Sevilla, A., Marquez, L., Nuez-Ortín, W., Moyano, F.J., 2013. Effect of partially protected butyrate used as feed additive on growth and intestinal metabolism in sea bream (*Sparus aurata*). Fish Physiology and Biochemistry 39, 1567-1580.

Said, H.M., Mohammed, Z.M., 2006. Intestinal absorption of water-soluble vitamins: an update. Curr. Opin. Gastroenterol. 22, 140-146.

Salze, G.P., Davis, D.A., 2015. Taurine: a critical nutrient for future fish feeds. Aquaculture 437, 215-229.

Sasso, O., Pontis, S., Armirotti, A., Cardinali, G., Kovacs, D., Migliore, M., Summa, M., Moreno-Sanz, G., Picardo, M., Piomelli, D., 2016. Endogenous N-acyl taurines regulate skin wound healing, Proc. Natl. Acad. Sci. U. S. A. 113, E4397-E4406.

Simó-Mirabet, P., Felip Edo, A., Estensoro, I., Martos-Sitcha, J.A., De las Heras, V., Calduch-Giner, J., Puyalto, M., Karalazos, V., Sitjà-Bobadilla. A., Pérez-Sánchez, J. 2018. Impact of low fish meal and fish oil diets on the performance, sex steroid profile and male-female sex reversal of gilthead sea bream (*Sparus aurata*) over a three-year production cycle. Aquaculture 490, 64-74.

Simó-Mirabet, P., Piazzon, M.C., Calduch-Giner, J.A., Ortíz, A., Puyalto, M., Sitjà-Bobadilla, A., Pérez-Sánchez, J., 2017. Sodium salt medium-chain fatty acids and Bacillus-based probiotics strategies to improve growth and intestinal health of gilthead sea bream (*Sparus aurata*). PeerJ 5, e4001.

Stabler, S.P., Allen R.H., 2004. Vitamin B12 deficiency as a world-wide problem. Annu. Rev. Nutr. 24, 299-326.

Stiboller, M., Raber, G., Francesconi, K.A., 2015. Simultaneous determination of glycine betaine and arsenobetaine in biological samples by HPLC/ICPMS/ESMS and the application to some marine and freshwater fish samples. Microchem. J. 122, 172-175.

Tacon, A.G.J., Metian, M., 2008. Global overview on the use of fish meal and fish oil in industrially compounded aquafeeds: Trends and future prospects. Aquaculture 285, 146-158.

Tai, S.S.-C., Bedner, M., Phinney, K.W., 2010. Development of a candidate reference measurement procedure for the determination of 25-hydroxyvitamin D3 and 25-hydroxyvitamin D2 in human serum using isotope-dilution liquid chromatography–tandem mass spectrometry. Analytical Chemistry 82, 1942-1948.

Tocher, D.R., 2015. Omega-3 long-chain polyunsaturated fatty acids and aquaculture in perspective. Aquaculture 449, 94-107.Turchini, G.M., Hermon, K.M., Francis, D.S., 2018. Fatty acids and beyond: Fillet nutritional characterisation of rainbow trout (*Oncorhynchus mykiss*) fed different dietary oil sources. Aquaculture 491, 391-397.

Une, M., Goto, T., Kihira, K., Kuramoto, T., Hagiwara, K., Nakajima, T., Hoshita, T., 1991. Isolation and identification of bile salts conjugated with cysteinolic acid from bile of the red seabream, Pagrosomus major. J. Lipid Res. 32, 1619-1623.

Venter, L., Jansen van Rensburg, P., Loots, D.T., Vosloo, A., Lindeque, J.Z., 2016. Untargeted metabolite profiling of abalone using gas chromatography mass spectrometry. Food Anal. Methods 9, 1254-1261.

Watanabe, F., Nakano, Y., Tachikake, N., Saido, H., Tamura, Y., Yamanaka, H., 1991. Vitamin B-12 deficiency increases the specific activities of rat liver NADH- and NADPH-linked aquacobalamin reductase isozymes involved in coenzyme synthesis. J. Nutr. 121, 1948-1954.

Wiklund, S., Johansson, E., Sjöström, L., Mellerowicz, E.J., Edlund, U., Shockcor. J.P., Gottfries, J., Moritz. T., Trygg J., 2008. Visualization of GC/TOF-MS-based metabolomics data for identification of biochemically interesting compounds using OPLS class models. Analytical Chemistry 80, 115-122.

Wishart, D.S., Jewison, T., Guo, A.C., Wilson, M., Knox, C., Liu, Y., Djoumbou, Y., Mandal, R., Aziat, F., Dong, E., Bouatra, S., Sinelnikov, I., Arndt, D., Xia, J., Liu, P., Yallou, F., Bjorndahl, T., Perez-Pineiro, R., Eisner, R., Allen, F., Neveu, V., Greiner, R., Scalbert, A., 2013 HMDB 3.0--The Human Metabolome Database in 2013. Nucleic Acids Res. 41, D801-D807.

Wold, S., Sjöström, M., Eriksson, L., 2001. PLS-regression: a basic tool of chemometrics. Chemometrics Intelligent Lab. Syst. 58, 109–130.Yancey, P.H., 2005. Organic osmolytes as compatible, metabolic and counteracting cytoprotectants in high osmolarity and other stresses. J. Exp. Biol. 208, 2819-2830.

CAPÍTULO III

AUTENTIFICACIÓN DE ALIMENTOS

III.1. Artículo científico 3

Food Control 70 (2016) 350–359

Contents lists available at ScienceDirect

Food Control

journal homepage: www.elsevier.com/locate/foodcont

Metabolomic approach for Extra virgin olive oil origin discrimination making use of ultra-high performance liquid chromatography – Quadrupole time-of-flight mass spectrometry

Rubén Gil-Solsona, Montse Raro, Carlos Sales, Leticia Lacalle, Ramon Díaz, María Ibáñez, Joaquim Beltran, Juan Vicente Sancho*, Felix J. Hernández

Research Institute for Pesticides and Water (IUPA), Avda. Sos Baynat, s/n, University Jaume I, 1201 Castellón, Spain

ARTICLE INFO

Article history:
Received 12 March 2016
Received in revised form
14 May 2016
Accepted 9 June 2016
Available online 10 June 2016

Keywords:
Extra virgin olive oil
Food authenticity
Liquid chromatography
High resolution mass spectrometry
Quadrupole time-of-flight
Metabolomics

ABSTRACT

The fraudulent miss-description on food product labels regarding origin or composition is a widespread problem. In this work, a metabolomic approach based on the use of ultra-high performance liquid chromatography coupled to quadrupole time-of-flight mass spectrometry (UHPLC-QTOF-MS) has been applied to identify the differentiating chemical markers that allow geographic origin discrimination between different Spanish Extra Virgin Olive Oils (EVOOs). For this purpose, ninety EVOOs from 6 Spanish regions were analyzed. Data processing consisted on peak picking, retention time alignment and response normalization. Partial Least Square Discriminant Analysis (PLS-DA) and orthogonal PLS-DA (OPLS-DA) were applied to identify the most significant markers that allow groups separation. Twelve different compounds were found to correctly separate the EVOOs from their origin and 7 of them could be tentatively identified. The results of our work suggest that UHPLC-QTOF MS-based metabolomic analysis is a suitable approach for biomarker-detection in the food quality/safety field.

1. Introduction

The traditional Mediterranean nutrition is widely known around the world as a very healthy diet. It was founded around the Mediterranean Sea, which gave it the name. In this entire geographical zone, olive crop has a widespread culture and, therefore, olive oil has a central position (Fazio & Ricciardiello, 2014; Vasto et al., 2014). The quality of olive oil is a key feature to ensure its healthy characteristics as well as its organoleptic qualities. Olive oil can be classified in four major quality groups: *Extra-Virgin Olive Oil (EVOO)*, *Virgin Olive Oil (VOO)*, *Olive Oil* and *Olive-pomace Oil* (European regulations EEC 2568/91 and EU 1348/2013). The quality of olive oils is highly correlated with several mechanical treatments in both harvesting and production processes (Angerosa et al., 2004), with consequences in the quality and final prize of the product in the market. For this reason, the importance of ensuring the quality of EVOOs has been reflected in several studies (Aparicio, Morales, Aparicio-Ruiz, Tena, & García-González, 2013).

On the one hand, the requirements for EVOO authenticity and differentiation from less quality Olive Oils are of essential concern for today's society and industry. In this sense, the adulteration of olive oils with other vegetal oils (Aparicio & Aparicio-Ruiz, 2000) is a fraudulent activity that leads to lower-quality olive oil. A great effort has been made in order to evaluate and set their authenticity (Faria, Cunha, Paice, & Oliveira, 2010). Chromatographic techniques, both Gas Chromatography (GC) (Angerosa et al., 2004; Gamazo-Vázquez, Garci;;a-Falcón, & Simal-Gándara, 2003) and Liquid Chromatography (Galeano Diaz, Durán Merás, Sánchez Casas, & Alexandre Franco, 2005), coupled to Mass Spectrometry (MS) have been the most employed. Different alternative techniques have been also evaluated, as Fourier Transformed Infra-Red (FTIR) (Maggio, Cerretani, Chiavaro, Kaufman, & Bendini, 2010), Nuclear Magnetic Resonance (NMR) (Dais & Hatzakis, 2013) or even Direct Analysis in Real Time (DART) (Vaclavik, Cajka, Hrbek, & Hajslova, 2009) coupled to MS.

Not only the quality but also geographical origin is important to ensure the value and organoleptic properties of this kind of premium foodstuff. Spain is the first olive oil producer around the world with more than 40% of total production, and counts with the rest of Europe (mainly Greece and Italy) approximately 70% of total

* Corresponding author.
E-mail address: sanchoj@uji.es (J.V. Sancho).

http://dx.doi.org/10.1016/j.foodcont.2016.06.008
0956-7135/

Metabolomic approach for Extra virgin olive oil origin discrimination making use of ultra-high performance liquid chromatography - Quadrupole time-of-flight mass spectrometry

Ruben Gil-Solsona, Montse Raro, Carlos Sales, Leticia Lacalle, Ramon Díaz, María Ibáñez, Joaquim Beltran, Juan Vicente Sancho* , Felix J. Hernández

Research Institute for Pesticides and Water (IUPA), Avda. Sos Baynat, s/n, University Jaume I, 1201 Castellon, Spain

Abstract

The fraudulent miss-description on food product labels regarding origin or composition is a widespread problem. In this work, a metabolomics approach based on the use of ultra-high performance liquid chromatography coupled to quadrupole time-of-flight mass spectrometry (UHPLC-QTOF-MS) has been applied to identify the differentiating chemical markers that allow geographic origin discrimination between different Spanish Extra Virgin Olive Oils (EVOOs). For this purpose, ninety EVOOs from 6 Spanish regions were analyzed. Data processing consisted on peak picking, retention time alignment and response normalization. Partial Least Square Discriminant Analysis (PLS-DA) and orthogonal PLS-DA (OPLS-DA) were applied to identify the most significant markers that allow groups separation. Twelve different compounds were found to correctly separate the EVOOs from their origin and 7 of them could be tentatively identified. The results of our work suggest that UHPLC-QTOF MS-based metabolomics analysis is a suitable approach for biomarker-detection in the food quality/safety field.

Introduction

The traditional Mediterranean nutrition is widely known around the world as a very healthy diet. It was founded around the Mediterranean Sea, which gave it the name. In this entire geographical zone, olive crop has a widespread culture and, therefore, olive oil has a central position (*Fazio & Ricciardiello, 2014; Vasto et al., 2014*). The quality of olive oil is a key feature to ensure its healthy characteristics as well as its organoleptic qualities. Olive oil can be classified in four major quality

groups: *Extra Virgin Olive Oil* (EVOO), *Virgin Olive Oil* (VOO), *Olive Oil* and *Olive pomace Oil* (*European regulations EEC 2568/91 and EU 1348/2013*). The quality of olive oils is highly correlated with several mechanical treatments in both harvesting and production processes (*Angerosa et al., 2004*), with consequences in the quality and final prize of the product in the market. For this reason, the importance of ensuring the quality of EVOOs has been reflected in several studies (*Aparicio, Morales, Aparicio-Ruiz, Tena, & García-Gonzalez, 2013*).

On the one hand, the requirements for EVOO authenticity and differentiation from less quality Olive Oils are of essential concern for today's society and industry. In this sense, the adulteration of olive oils with other vegetal oils (*Aparicio & Aparicio-Ruíz, 2000*) is a fraudulent activity that leads to lower-quality olive oil. A great effort has been made in order to evaluate and set their authenticity (*Faria, Cunha, Paice, & Oliveira, 2010*). Chromatographic techniques, both Gas Chromatography (GC) (*Angerosa et al., 2004; Gamazo Vazquez, García-Falcon & Simal-Gandara, 2003*) and Liquid Chromatography (*Galeano Diaz, Duran Merás, Sánchez Casas, & Alexandre Franco, 2005*), coupled to Mass Spectrometry (MS) have been the most employed. Different alternative techniques have been also evaluated, as Fourier Transformed Infra-Red (FTIR) (*Maggio, Cerretani, Chiavaro, Kaufman, & Bendini, 2010*), Nuclear Magnetic Resonance (NMR) (*Dais & Hatzakis, 2013*) or even Direct Analysis in Real Time (DART) (*Vaclavik, Cajka, Hrbek, & Hajslova, 2009*) coupled to MS.

Not only the quality but also geographical origin is important to ensure the value and organoleptic properties of this kind of premium foodstuff. Spain is the first olive oil producer around the world with more than 40% of total production, and counts with the rest of Europe (mainly Greece and Italy) approximately 70% of total worldwide production (http://www.internationaloliveoil.org/estaticos/view/131-world-olive-oil-figures?lang¼es_ES, last accessed 22/02/2016). Geographical discrimination plays an important role in issue, where analytical chemistry provides advanced tools to ensure the origin of olive oils. Different studies have been developed with olive oils from Italy (*Portarena, Gavrichkova, Lauteri, & Brugnoli, 2014*), Greece (*Longobardi et al., 2012)*, France (*Cavalli, Fernandez, LizzaniCuvelier, & Loiseau, 2004*) or Tunisia (*Camin et al., 2016*) in order to differentiate oils from different countries highlighting the relevance of these studies in actual research perspectives.

Moreover, designation of origin is also important for consumers (*Erraach, Sayadi, Gomez & Parra-Lopez, 2014*). For this reason, Protected Designations of Origin (PDO) Regulatory Council was created in order to ensure the right identification of different production zones of EVOOs (like Spain, Italy, Greece, France, etc). Regarding Spain, different studies have been found in the literature differentiating spanish olive oil varieties (*Vergara-Barberan, Lerma-García, Herrero-Martínez, & Simo-Alfonso, 2015*) as well as PDOs (*Beltran, Sánchez-Astudillo, Aparicio & García-González, 2015; García-González, Luna, Morales & Aparicio, 2009*). However, to the best of our knowledge, a full-Spanish geographical characterization study has not been published in the literature yet.

In order to achieve this goal, different techniques have been raising up in the last decade, being metabolomics the-state-of-the-art in this field. This recent approach has been increasingly used over the last years, based on the discovery of unknown, statistically significant compounds in the matrix, which allow to differentiate between classes (*Cevallos-cevallos, Etxeberria, Danyluk, & Rodrick, 2009*). Metabolomics has been employed for quality food control, in matrices such as wine (*Ali, Maltese, Toepfer, Choi, & Verpoorte, 2011*) or oranges (*Díaz, Pozo, Sancho, & Hernandez, 2014*).

Metabolomics appeared, firstly, as a nuclear magnetic resonance (NMR)-based technique. NMR was initially employed because of its universality and versatility as well as robustness; however, the low sensitivity and the high cost of this technique are important limitations, in addition to the higher sample quantity commonly required for the analysis. These drawbacks can be solved by using high resolution MS (HRMS) techniques coupled to both GC/LC, which appears nowadays as one of the most efficient approaches in the field of metabolomics. Different strategies have been employed, mainly GC and LC coupled to MS (*Gallart-Ayala, Chereau, Dervilly- Pinel, & Le Bizec, 2015*), but multiplatform metabolomics approaches are also emerging, for example combining GC-LC-CE (*Rojo et al., 2015*).

GC-MS has been the most widely used technique for olive oil characterization to investigate volatile and semi volatile compounds. In most cases, isolation of the analytes by means of an extraction step has to be carried out before chromatographic determination using purge-trap systems or HS-SPME as well as by direct HS sampling and injection (*Angerosa et al., 2004; Flath, Forrey, & Guadagni, 1973; Hu et al., 2014; Jimenez, Aguilera, Beltran, & Uceda, 2006; Pouliarekou et al., 2011*) or even

GCxGC applications (*Peres et al., 2013; Purcaro, Cordero, Liberto, Bicchi, & Conte, 2014*). LC is surely the best complement to directly analyze less volatile compounds in the matrix, as it requires less sample treatment and reduces compound losses, as it has been reported for essential oils (*Do, Hadji-Minaglou, Antoniotti, & Fernandez, 2015*).

The aim of this study was to discover and identify relevant biomarkers that allow classifying different Spanish EVOOs based on their geographical origin using Ultra-high Performance Liquid chromatography (UHPLC) coupled to Quadrupole Time-of-Flight Mass Spectrometry (QTOF MS), in combination with multivariate analysis (PLS-DA, OPLS-DA).

Materials & methods

Chemicals and reagents

HPLC-grade water was obtained by purifying demineralized water in a Mili-Q plus system from Millipore (Bedford, MA, USA). HPLC-grade methanol (MeOH), HPLC-supergradient acetonitrile (ACN), HPLC-grade 1-butanol (BuOH), HPLC-grade 2-propanol, sodium hydroxide (NaOH, >99%) and reagent-grade ammonium acetate (NH4Ac) were obtained from Scharlab (Barcelona, Spain). Leucine-enkephalin and formic acid (HCOOH, 98-100%) were purchased from Sigma-Aldrich (Augsburg, Germany). Oleic acid (reagent-grade, 99%), palmitic acid (free acid Sigma-grade), linoleic acid (free, 99%), glyceryl trioleate (Sigma-grade, 99%), Diolein and 1-Monooleoyl-RAC-Glycerol were also acquired from SigmaAldrich.

Samples

One of the most important steps in metabolomics is sampling and sample characterization. Samples used for model construction should have a complete traceability in order to obtain significant and valid results through sample classes analyzed. For this reason, sampling process was designed in collaboration with InterCoop, the Valencian Community olive oil cooperative, which collects and distributes the highest part of Valencian produced olive oil. Furthermore, Spanish cooperatives are associated with the Spanish Agriculture Ministry in order to promote the high

quality of Spanish EVOOs. This relationship was established through the "*Patrimonio Comunal Olivarero*" Foundation (http://www.pco.es/default.aspx) that promotes and distributes EVOOs from all the Spanish regions with the required traceability.

In this sense, 57 EVOOs from different cultivar zones of Spain (see **Fig. S1**) were acquired in *Patrimonio Comunal Olivarero* (Madrid, Spain): 5 from Bajo Aragon, 10 from Cataluña (5 from *Tarragona* and 5 from *Girona*), 5 from Toledo, 8 from Navarra-Rioja (4 from *Navarra* and 4 from *La Rioja*) and 29 from Andalucía (4 from *Jaén*, 4 from *Sevilla*, 3 from *Granada*, 4 from *Sierra Segura*, 4 from *Sierra Mágina*, 3 from *Málaga*, 4 from *Almeria* and 3 from *Córdoba*). Additionally, 33 Extra Virgin Olive Oil from Valencian Community (CV) were provided by InterCoop (*Castellon, Spain*), including: 9 from *Maestrat*, 6 from *La Plana Alta i Alcalaten*, 7 from *Serra Espada i Calderona*, 8 from *Serrania del Túria* i la *Ribera del Magro*, 1 from *Vinalopó* and 2 from *Utiel-Requena*.

These samples were chosen to achieve a good representation of all the Spanish Olive cultivar zones regarding their geographical distribution. Test samples were acquired in the same specialty store (*Patrimonio Comunal Olivarero, Madrid, Spain*) for model testing in a different season: 2 from Bajo Aragón, 2 from Cataluña, 2 from Toledo, 2 from Valencian Community, 1 from Rioja, 1 from Navarra and 5 from Andalucía (1 from *Granada*, 1 from *Córdoba*, 1 from *Málaga*, 1 from *Sevilla* and 1 from *Jaén*). In total, fifteen different samples were purchased, being representative samples of all Spanish EVOOs.

Sample treatment

Two different sample treatments were applied to the olive oils, which were stored at room temperature. For the polar components (polar fraction), a liquid-liquid extraction (LLE) was carried out mixing 1 mL of EVOO sample with 1 mL of methanol. 0.75 mL of the supernatant was taken and dried using a MiVac Duo concentrator.

The residue was reconstituted with 0.75 mL of H_2O:MeOH (1:1, v/v). After stirring, 70 mL of each sample (approximately 10% of the sample) were pooled to obtain a Quality Control (QC), which was injected at the beginning of the sample batch for column conditioning and, then, every 10

samples to control possible undesired instrument drifts and to correct data normalization. The rest was frozen at -24 ºC until sample analysis.

For the less polar compounds (non-polar fraction), 150 mL of each EVOO were ten-fold diluted with 1.50 mL of BuOH and stirred for 30 s, pooling 200 mL of each sample to obtain the QC solution. Again, samples were frozen at -24 ºC until analysis.

Instrumentation

A Waters Acquity UPLC system (Waters, Milford, MA, USA) was interfaced to a hybrid quadrupole-TOF high resolution (HRMS) mass spectrometer (Xevo G2 QTOF, Waters, Manchester, UK) using a Z-spray-ESI interface operating in both positive and negative ionization modes with resolution of the TOF mass spectrometer about 20,000 at full width half maximum (FWHM). The UHPLC separation was performed using a 100x2.1 mm Acquity UPLC BEH C18 (1.7 mm particle size) analytical column (Waters) at 300 mL/min (For further details see Supplementary Material).

Data processing

The UHPLC-(Q)TOF MS data were converted from proprietary (.raw, Waters) to generic (.cdf, NetCDF) format using Databridge application (within MassLynx v 4.1; Waters Corporation) and preprocessed using XCMS free R package (*Smith, Want, O'Maille, Abagyan, & Siuzdak, 2006*). For peak picking, centWave feature selection algorithm was employed, considering peak width ranging from 5 to 20 s, with at least 3 scans above 1000 counts, a signal to noise ratio of 10 and 15 ppm mass tolerance. Peak grouping (bandwidth from 10 down to 0.5 s) was performed to match detected features across samples before peak alignment step using the retcor() function. In order to create a list with all these compounds as well as to obtain peak areas, the function fillPeaks() was used. Peaks are labeled as MxxxTyyy, with xxx referring to its nominal mass and yyy to the corrected retention time in seconds. A Loess normalization method was applied to the samples, in order to correct instrumental drift.

Multivariate analysis was carried out using EZInfo (*Umetrics, Sweden*). First of all, Principal Component Analysis (PCA) was applied in order to check the correct normalization of the samples by analyzing the behavior of the QC samples, ideally clustered in the center of the Score Plot. This

type of analysis is employed for dimension reduction; as it is not guided, the QC samples joining in the center of the plot implies the correct normalization of the batch.

After that, Partial Least Square-Discriminant Analysis (PLS-DA) was used to distinguish samples according their geographical origin. This analysis creates orthogonally directed components in order to explain the preselected groups. If good separations are observed we can have an idea about the quantity of "good markers" that should be expected for group separation. In our case, preliminary results showed group separation. The final groups were Andalucía, Cataluña, Rioja-Navarra, Toledo, Comunidad Valenciana (CV) and Aragon.

Orthogonal PLS-DA (OPLS-DA) was then applied, by facing each group to the rest, in order to obtain the most relevant markers for each group. The workflow can be seen in **Fig. 1**. After performing OPLS-DA for each of the six groups, a list of biomarkers for the separation of one group against the rest was obtained.

The final objective of these statistical methods is to obtain, by an untargeted approach, a reduced list of compounds to develop a targeted method for origin control.

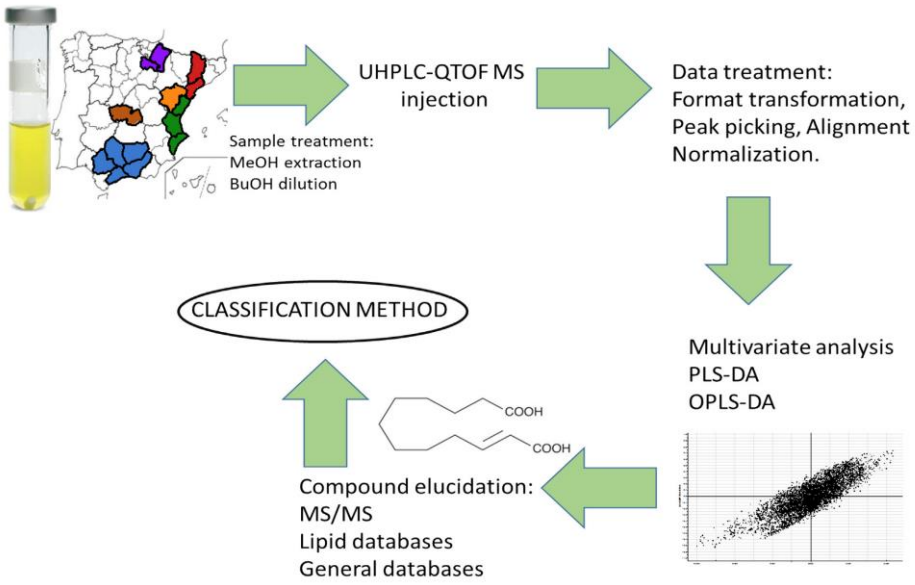


Fig. 1: Data treatment processing followed during the metabolomics workflow

Elucidation workflow

The following workflow for candidate elucidation was proposed. For significant markers (MxxxTyyy), the accurate mass was retrieved from the XCMS final table and LE and HE mass spectra were extracted from raw data to visualize fragments. From the accurate mass, molecular formula/s were derived using the built-in elemental composition calculator. Then, these formulas as well as accurate masses were searched for possible candidates in specific databases as Lipidmaps (http://www.lipidmaps.org/data/structure/, last accessed 22/02/2016) or lipidbank (http://lipidbank.jp/, last accessed 22/02/2016). Later MS/MS experiments were carried out for proper fragment assignment in HE spectrum or for proper isolation of the relevant precursor ion if coeluting compounds render a high complex HE spectrum. For potential candidates, if one or more molecules were obtained, MassFragment software provided by MassLynx (*Waters Corporation*) was used to discard candidates based on feasible explanation of the fragments/product ions.

When no properly explained candidate was found for a specific marker, feasible elemental compositions for (de)protonated molecule (and/or adducts) as well as fragments were calculated. The molecular formula was introduced then in general chemical databases as ChemSpider (http://www.chemspider.com/, last accesed 22/02/2016) for additional candidates searching. Tentative explanation of the fragmentation was also carried out for these molecules by in-silico fragmentation softwares as MetFrag (http://msbi. ipb-halle.de/MetFrag/, last accessed 22/02/2016). Moreover, when available, the measured MS/MS spectra were also compared to free mass spectral databases such as MassBank (http://massbank.ufz. de/MassBank/, last accessed 22/02/2016) or Metlin (https:// metlin.scripps.edu/index.php, last accessed 22/02/2016). Finally, when too many or none potential candidates were found, the marker was simply annotated using its XCMS label (**Fig. S2**).

Prediction model

In order to predict the geographical region of the different EVOOs, the peak area of the finally selected markers was integrated from the raw data with TargetLynx (quantitative application

manager from MassLynx) and log2 area transformed in the 90 model samples. Afterwards, data was transferred to SPSS Statistics 22 (*IBM corporation*) building a prediction model for EVOO classification. A stepwise variable selection model was employed (Wilks lambda for testing the equality of group means, probability of F:(Entry ¼ 0.05, Removal ¼ 0.10)) for the given peaks. This model was employed to ensure that all the selected markers included in the method (which provides information about the goodness of these markers) are used to create the model. The method employed to create the discrimination model was later tested with 15 additional EVOO second season samples.

For normalization, the integrated areas of markers (using TargetLynx) were corrected using the relation between mean values for each marker in both sample batches (model and testing samples). The correction and utility of this kind of normalization will be discussed in Section 3.4.

Results and discussion

Chromatographic conditions

Chromatographic conditions are essential in metabolomics workflow in order to obtain good separation that usually provides better results. Typically, a set of known compounds is used to obtain the best chromatographic conditions. However, in untargeted metabolomics, dealing with unknowns, specific chromatographic requirements cannot be optimized for each experiment. Therefore, in this work the Based Peak Ion (BPI) chromatogram was evaluated to achieve the maximum peak distribution during preliminary experiments. Olive oil is a non-polar matrix, therefore a reversed phase Acquity BEH C18 column (100x2.1 mm, 1.7 mm particle size) was chosen for the analysis of both non-polar and polar fractions. Both positive and negative ionization modes were acquired in order to maximize the information obtained about the samples to facilitate a successful separation between groups.

The mobile phases for the polar fraction were water/methanol adding formic acid for better peak shape, as in previous foodomics studies carried out in our laboratory (*Díaz et al., 2014*), which showed narrow and resolved peaks. The gradient was linear from 10 to 90% methanol in 14

min to separate the maximum number of compounds of the sample. From 14 to 20 min the percentage of organic modifier was increased up to 100% to elute non-polar compounds, which are the most retained under reversed-phase separations. With this elution gradient (see Section 2.4) more than 5800 *m/z*_RT pairs per sample were obtained in positive ionization and about 1500 in negative ionization mode when performing peak picking with XCMS software.

For a better separation and resolution of the peaks in the nonpolar fraction, preliminary experiments were carried out with a mix of several free fatty acids, mono-, di- and triacylglycerides prepared in both n-butanol and 2-propanol solvents. Different mobile phases were investigated including acetonitrile:2-propanol and acetonitrile:*n*-butanol. The best chromatographic peak shape and distribution was observed using n-butanol as strong organic modifier. The addition of water to the weaker mobile phase A was also studied. The miscibility of n-butanol with water was posible due to the presence of acetonitrile. The addition of a small quantity of water helped in the separation of free fatty acids. Finally, the addition of 15% water to acetonitrile was selected, as it improved the peak shape for the more "polar" compounds (free fatty acids and monoacylglycerides) while nonpolar compounds (di- and triacylglycerides) were better resolved without affecting the elution time.

The hydrophobicity of olive oil compounds caused that a small amount of these components remained on the injection valve, affecting the following injection. To avoid, or at least minimize this problem, a further no-injection run was performed after each sample injection with a higher initial percentage (50%) of strong modifier (n-BuOH) for cleaning out the injection port before the next sample.

Data processing

Due to the apparent similarity of BPI chromatograms for EVOO of different regions, as shown in **Fig. S3**, data analysis software seems mandatory for markers discovery. A first view of these data showed 5-12 s wide chromatographic peaks with up to 106 counts in the case of clearly visible peaks. For detection of smaller components buried into the TIC background, a minimum peak height of 1000 counts was selected.

For peak picking process, XCMS software was employed. This package, freely available on the internet (http://www.bioconductor.org/packages/release/bioc/html/xcms.html last accessed 22/02/ 2016), contains specific functions to extract as much information as possible from samples (see Section 2.5) and helps to correct potential retention time deviations across sample batch analysis (118 injections/47 h in our case).

After XCMS processing, 5896 and 1544 features (*m/z* ions) were obtained from the polar fraction in positive and negative ionization modes, respectively. Regarding the non-polar fraction, 2457 (ESIþ) and 312 (ESI) features were found.

For peak area normalization, a mean centering normalization process was applied. As shown in **Fig. S4**, sample mean intensity of non-normalized data was slightly dropping along the batch. This drift was corrected after applying the normalization process. This correction can be better observed for QCs, which were injected every 10 samples. A QC signal dropout is clearly observed on the raw data but completely corrected after normalization. In the next step, statistical analysis was carried out.

In the next step, statistical analysis was carried out. Principal Component Analysis was initially performed to check the normalization quality based on the QCs grouping as well as to investigate possible undesired or unexpected outliers on the samples. Then, Partial Least Squares-Discriminant Analysis (PLS-DA) was performed to better observe group separation. This first data visualizing showed that 93.1% of total variance was explained by the first 12 components (**Fig. S5b**). As can be seen in **Fig. S5a** all the QC samples (pink squares) appear at the center, whereas the different EVOO samples appear correctly grouped in the Score Plot. According to these data, it seems feasible to separate the samples using this UHPLC-HRMS method.

In order to highlight the best features for EVOOs differentiation according to their origin, an Orthogonal PLS-DA (OPLS-DA) was finally employed. Each group was faced to the rest, highlighting the most important compounds in all cases. As an example, **Fig. 2**, corresponding to OPLS-DA analysis for non-polar positive analysis, shows all the compounds annotated across all samples, when Andalucia EVOOs are compared against the rest. In this case, as can be seen in **Table 1** marked as "Andalucia vs rest", two markers (M880T325 and M678T381), highlighted in the S-Plot,

were selected for building the classification model and subsequent elucidation. This process was repeated for the rest of the regions, selecting finally twelve markers for origin discrimination listed in **Table 1.**

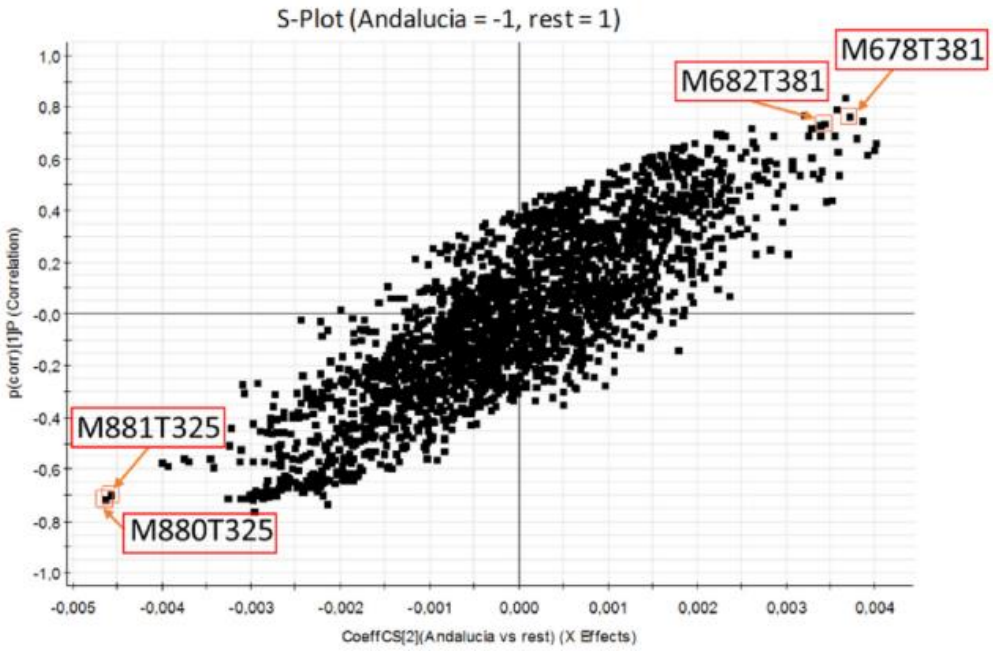


Fig. 2: S-Plot of OPLS-DA model for Andalucía vs the rest of the regions. M880T325 and M881T325 corresponde to compound #9 (see Table 1) and M678T381 and M682T381 to compound #4.

These selected compounds along with their peak areas, integrated by TargetLynx (MassLynx, Waters) software, were log2 transformed and introduced into SPSS Statistics 22 (IBM) in order to obtain the discrimintating power of the Stepwise model (details in section 2.7) for these samples. The statistical method created 5 different functions, which allowed to correctly classify 94.4% of all samples after cross-validation of original cases for both integration modes as can be seen on **Table 2** and **Fig. 3**. In this figure, FUNCTION 1, FUNCTION 2 and FUNCTION 3 are represented as X-axis, Y-axis and Z-axis respectively, but FUNCTION 4 and FUNTION 5 were also employed for each sample to classify them. As can be observed, the first three components allows to distinguish two different groups: Andalucia, Comunidad Valenciana and Toledo in the upper part, and Aragon, Cataluña.

Table 1: LC-QTOF MS accurate mass measurement for selected compounds for origin EVOO separation.

RT (min)	Compound	Formula $[M+H]^+$	Experimental *m/z*	Theoretical *m/z*	Error (mDa/ppm)	Aliquot	Regions confronted	Fragment ions at high energy
9.87	Traumatic acid	$C_{12}H_{20}O_4$	229.1446	229.1440	0.6/2.6	NS+	Aragón vs rest	211.134 $[M+H-H_2O]^+$ 183.1373 $[M+H-C_2H_6O]^+$
10.59	5-O-Methyllatofolin	$C_{18}H_{20}O_4$	301.1440	301.1440	0.0/0.0	NS+	Aragón vs rest	123.1170 $[M+H-HCOOH-CH_3COOH]^+$ 131.0497 $[M+H-C_9H_{14}O_3]^+$
7.52	Oleoyl, Linoleoyl, Palmitoyl-glycerol	$C_{55}H_{98}O_6$	855.7464	855.7442	2.2/2.6	LI+	CV vs rest	599.5037 $[M+H-C_{16}H_{32}O_2]^+$ 573.4879 $[M+H-C_{18}H_{34}O_2]^+$ 577.5254 $[M+H-C_{18}H_{32}O_2]^+$
6.35	Oleoyl, Linoleoyl. Acetyl-glycerol	$C_{41}H_{72}O_6$	661.5419	661.5407	1.2/1.8	LI+	Andalucía vs rest	379.2860 $[M+H-C_{18}H_{34}O_2]^+$ 381.3009 $[M+H-C_{18}H_{32}O_2]^+$
1.88	9,13-dihydroxy-11-octadecenoic acid	$C_{18}H_{34}O_4$	315.2531	315.2535	-0.4/-1.2	LI+	Aragón vs rest	263.2387 $[C_{18}H_{32}O_2 -H_2O+H]+$ 185.1187 $[M+H-C_8H_{18}O]^+$ 139.1128 $[M+H-C_9H_{20}O_3]^+$ 121.1016 $[M+H-C_{19}H_{22}O_4]^+$
3.82	11-ethyl-1,25-dihydroxyvitamin D3	$C_{29}H_{48}O_3$	445.3680	445.3682	-0.2/-0.4	LI+	Navarra-Rioja vs rest	293.2528 $[M+H-C_9H_{312}O_2]^+$ 165.0916 $[M+H-C_{19}H_{36}O]^+$
11.66	14-oxo-pentadecanoic acid	$C_{15}H_{28}O_3$	257.2113	257.2117	-0.4/-1.6	NS+	CV vs rest	239.2012 $[M+H-H_2O]^+$ 225.1852 $[M+H-CH_4O]^+$ 207.1744 $[M+H-CH_6O_2]^+$
9.97	M323T599	$C_{17}H_{22}O_6$	323.1491	323.1495	-0.4/-1.2	NS-	Cataluña vs rest	187.0969 $[M-H-C_8H_8O]^-$ 115.0421 $[M-H-C_{12}H_{16}O_3]^-$ 97.0653 $[M-H-C_{11}H_{14}O_5]^-$
5.41	M880T325	$C_{49}H_{83}O_{13}$	879.5836	879.5834	0.2/0.2	LI+	Andalucía and Toledo vs rest	759.5380 $[M+H-2{\cdot}C_2H_4O_2]^+$ 699.5240 $[M+H-3{\cdot}C_2H_4O_2]^+$ 639.4970 $[M+H-4{\cdot}C_2H_4O_2]^+$
11.67	M687T701	$C_{36}H_{47}O_{13}$	687.3026	687.3017	0.9/1.3	NS-	Andalucía vs rest	293.1401 $[M-H-C_{20}H_{28}O_8]^-$ 231.1378 $[M-H-C_{21}H_{28}O_{11}]^-$
14.76	M501T889	-	501.3928	-	-	NS+	Cataluña vs rest	339.2918 $[M-H-254.2001]^+$
6.07	M593T364	-	593.4919	-	-	LI+	Toledo vs rest	313.2773$[M-H-280.2325]^+$

For the aliquot notation, NS correspons to polar aliquot and LI corresponds to non-polar fraction. While symbols + and – indicate the polarity of the ES

and Navarra-Rioja in the lower zone. Center plot (with Andalucia, Comunidad Valenciana and Toledo groups only represented) shows the goodness of the method with the first three functions that allow to separate these samples, as cannot be properly appreciated in the left plot. Finally, right plot shows the separation of Aragon, Cataluña and Navarra-Rioja. These three functions show a tolerable separation, but the combination of these 5 functions allows the method to separate correctly the groups among the rest at 94.4%. 3-D plots have been drawn with plotly R package (https://plot.ly/plot#, last accessed 22/ 02/2016).

Table 2: SPSS Classification method results for each.

Pronosticated group								
	Region	*ANDALUCIA*	*ARAGON*	*COMUNIDAD VALENCIANA*	*CATALUÑA*	*RIOJA NAVARRA*	*TOLEDO*	Total
NUMBER	*ANDALUCIA*	28	-	-	1	-	-	29
	ARAGON	-	5	-	-	-	-	5
	C.V.	1	1	31	-	-	-	33
	CATALUÑA	-	-	-	10	-	-	10
	RIO.-NAV.	-	1	-	-	7	-	8
	TOLEDO	1	-	-	-	-	4	5

In each line, cell number in bold corresponds to the number of cases correctly classified in different groups

TargetLynx integration process was finally preferred, as this approach was easier and faster to implement in the next analysis (model testing) and further analysis of EVOO samples. Only new rawdata and previously created integration method for selected markers are needed, in front of time-consuming format conversion, data processing and database searching for the selected markers for integrating them. Moreover, the possibility of different labeling during XCMS processing in different batches due to small variations in retention time (for example, M880T326 instead of M880T325) could lead to obtain no response for M880T326 during peak picking, incorrectly classifying the simple.

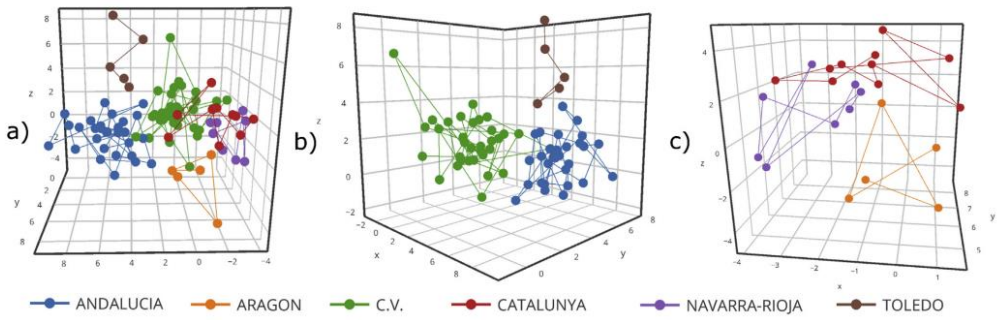


Fig. 3: Classification method for EVOO samples with SPSS Statistics. Function 1 and 2 are represented in X-axis and Y-axis. Z-axis corresponds to function 3 in the three plots: a) All groups representes, b)Toledo, Andalucia and Comunidad Valenciana and c) Cataluña, Navarra-Rioja and Aragon

Structural elucidation of candidate markers

Twelve selected markers were identified to provide a good separation between the different EVOOs based on their origin. Accurate mass was provided by XCMS for features revealed by OPLS-DA and a candidate structure was tried to assign by searching the accurate mass in the databases (see **Table 1**). Five out of the twelve markers could not be properly identified (compounds 8 – 12).

Compound 8 obtained from the analysis of the polar extract in negative ionization mode, was not found in specific databases, so accurate mass was searched in generic databases. Taking into account the accurate mass of observed product ions in MS^E experiments, only one molecular formula ($C_{17}H_{22}O_6$) could explained all the observed fragments but this formula could correspond to several candidates. In order to discard potential candidates, on-line in-silico fragmentation software like MetFrag was used but still several candidate structures remained plausible. Thus, this compound was named based on XCMS annotation. In the same way, compounds 9-12 were tried to be elucidated using information obtained from both specific and general databases. Finally compounds

9 and 10 were assigned to a single molecular formula but for compound 11 and 12 it was not even possible to decreased down to a unique formula and accurate mass was only given (see **Fig. S6**).

The remaining seven markers were tentatively elucidated based on their accurate mass spectra (LE, HE and MS/MS experiments) and database search (**Table 1**). As an illustrative example of the benefits of HRMS/MS combined with chemical and mass spectral databases, the elucidation of compound 1, found in the polar positive analysis, is shown below.

The LE (MS^E) QTOF MS spectrum of the marker 1 showed the presence of a protonated molecule $[M+H]^+$ at *m/z* 229.1438 and its sodium adduct $[M+Na]^+$ at *m/z* 251.1262. Several fragment ions were already observed in LE at *m/z* 211.1341, 183.1384, 123.1174 and 81.0697. Regarding HE data, the fragment ions at *m/z* 123.1173 and 81.0704 were still observed (**Fig. 4a**). In all cases, mass errors were lower than 2 mDa for the suggested protonated molecule, which corresponded to an empirical formula of $C_{12}H_{20}O_4$. In this case, additional MS/MS experiments could not be carried out due to the low intensity of the protonated molecule. However, when comparing the LE and HE spectra with Metlin online spectra (http:// metlin.scripps.edu/metabo_info.php?molid¼5882, last access 22/ 02/2016) a single hit was returned (**Fig. 4b**), with all the fragments being observed. In this way, marker 1 was tentatively identified as traumatic acid (see **Fig. 4c**).

This compound appears in Metlin database in both ionization modes. However, we have not seen it under negative ionization mode. The low sensitivity of our ElectroSpray Ionization (ESI) source in negative ionization mode compared to positive one in addition to the low response of this compound could lead its signal to disappear. On the other hand, the high similarity in positive mode led us to tentatively elucidate it. The ultimate confirmation would require the purchase of the reference standard, which at the time of writing this paper was not available in our laboratory.

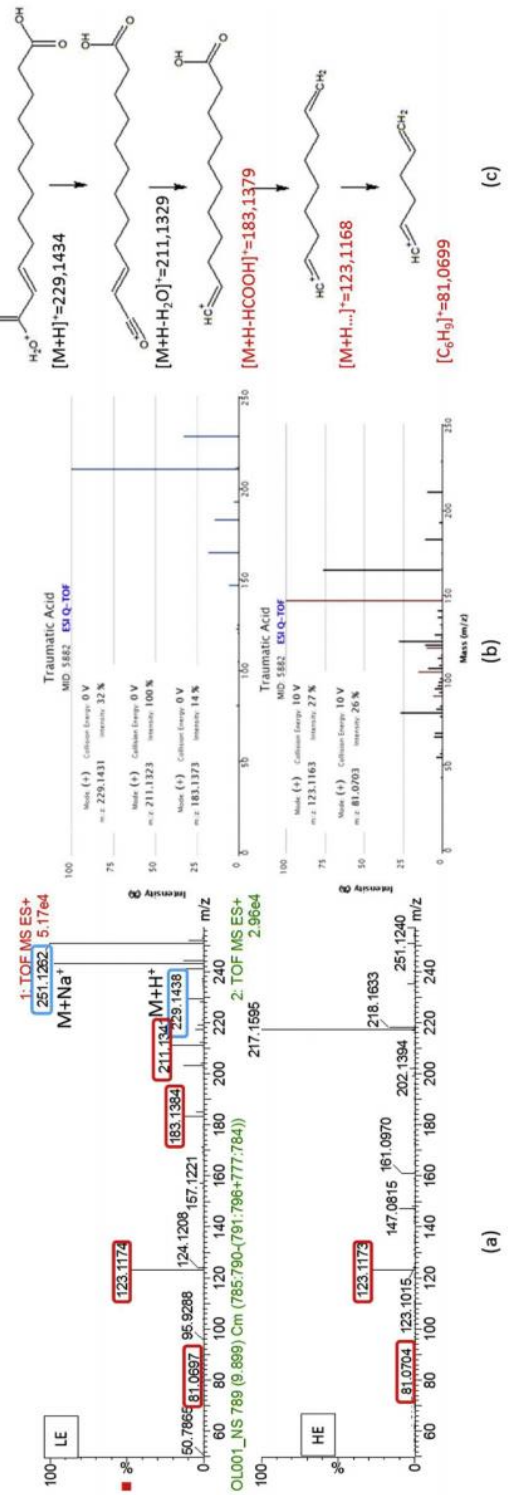


Fig. 4: Elucidation of marker #1 M251T599 from ESI+ in the polar aliquot: (a) LE spectrum (top), HE spectrum (bottom). Protonated molecule and sodium adduct are emphasized in blue, and fragments in red; (b) Uploaded Metlin MS/MS spectra for traumatic acid at 0 and 10 eV collision energy; (c) Traumatic acid fragmentation explained with the observed *m/z* peaks.

Another example is shown in **Fig. 5**, which illustrates the detection and identification of marker 3, a compound that was found as significant in the ESI+ non-polar fraction. As can be seen in the LE spectrum (**Fig. 5a**), based on the accurate mass difference of the base peaks at *m/z* 872 and 877 one could assign them to both ammonium and sodium adducts of a neutral molecule with MW 854 and an expected protonated exact mass of *m/z* 855.7411 (low intensity in the LE spectrum). In the HE spectrum (data not shown) only *m/z* 877 is still observed due to the high stability and poor fragmentation of sodium adducts, but different cluster ions around *m/z* 500-600 can now be observed, and the same results was obtained in tandem MS experiment (**Fig. 5b**). From *m/z* 877.7303 (**Fig. 5a**), the most plausible empirical formula for the sodium adduct was $C_{55}H_{98}O_6Na$, with 4.2 mDa (4.8 ppm) mass error. As this marker was found in the non-polar fraction and it shows 6 oxygen atoms, the most plausible candidate was a triacylglyceride. The fragment ions from MS/MS experiments at *m/z* 500e600 (*m/z* 599.5016($[M+H - Palmitic\ acid]^+$), 573.4872 ($[M+H - Oleic\ acid]^+$) and 577.5194 ($[M+H- Linolenic\ acid]^+$)) with mass errors lower than 2.5 mDa could be assigned to neutral losses of the different fatty acids present in this triglyceride. Another cluster ion can be observed around *m/z* 200-300 containing ions at *m/z* 265.2529 ($Oleic\ acid+H^+-H_2O$, +0.3 mDa) and 239.2377 ($Palmitic\ acid+H^+- H_2O$, +0.8 mDa). These peaks correspond to the fragmentation of the triglyceride with the charge retained on the leaving fatty acid. With all data available, this marker was tentatively identified as the triglyceride OLnP, following previously reported nomenclature (*Lísa & Holcapek, 2008*).

In a similar way, a total of 7 compounds were tentatively identified, including triacylglycerides, unsaturated acids or D3 Vitamin related compounds. For unequivocal confirmation of their identity, reference standards should be acquired and injected under the same conditions for retention time and fragmentation pathway evaluation. Thus, in a later stage target quantitative analysis directed towards these compounds would be feasible using simpler techniques as LC-MS/MS with triple quadrupole.

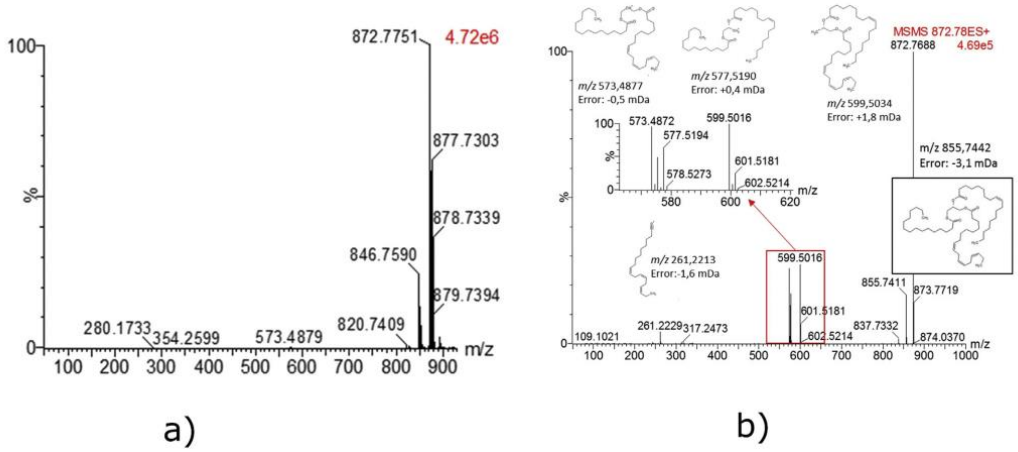


Fig. 5: Elucidation of marker #3 M873T451 from ESI+ in the non-polar aliquot: (a) LE spectrum and (b) MS/MS spectrum at 20 eV

Testing process

In order to evaluate the robustness of the prediction model, a testing process was performed using the responses of the twelve selected markers. For this purpose 15 additional EVOO samples were acquired from PCO, prepared, injected under the same conditions and marker responses (peak areas) were obtained. Normalized and non-normalized data were processed after log2 transformation. Ten of the fifteen samples were correctly classified with non-normalized data (66%). The remaining five samples were misassigned maybe as the result of using absolute non-normalized areas, due to possible differences in instrument response along the time, despite the use of log2 transformation. These differences were reduced after the application of a "normalization factor" (mean response of marker X in model samples/mean response of marker X in test samples), increasing up to 13/15 correct assignations (87%) even with six different classes and using samples from different seasoning (see **Table 3**). This "normalization factor" makes different sample batches comparable between them, without need of standard injections, not available for all the compounds. In this sense, the model prediction success is considered satisfactory for Spanish EVOO classification.

Table 3: Testing process results in number of samples. Correctly classified samples are in bold.

Pronosticated group								
	Region	*ANDALUCIA*	*ARAGON*	*COMUNIDAD VALENCIANA*	*CATALUÑA*	*RIOJA NAVARRA*	*TOLEDO*	Total
BELONGING	*ANDALUCIA*	4	-	-	-	-	1	5
GROUP	*ARAGON*	-	2	-	-	-	-	2
(norm. data)	*C.V.*	-	-	2	-	-	-	2
	CATALUÑA	1	-	-	1	-	-	2
	RIO.-NAV.	-	-	-	-	2	-	2
	TOLEDO	-	-	-	-	-	2	2

Conclusions

Spanish EVOO samples produced in different areas of Spain were analyzed by UHPLC QTOF-MS using a non-targeted metabolomic approach with the aim of discriminating the samples based on their geographic origin. The dataset was subjected to a first pretreatment with XCMS software and a subsequent multivariate statistical analysis. Pre-treatment of data worked properly, obtaining a time alignment along the samples of less than 1 s, and a successful total intensity correction along the batch. The OPLS-DA analysis highlighted some compounds that might be used as EVOO's origin markers, and 12 of them were used to group samples after statistical classification with SPSS Statistics software. These compounds help to predict the Spanish region of an EVOO sample, with a simple analysis of their peak areas. Several compounds were tentatively identified as triglycerides, Vitamin D3 related compounds or different organic acids. In order to validate the model, the method was applied to fifteen Spanish EVOO samples collected in a different season, and it was able to assign around 90% of them. Promising data have been obtained in this work to direct future research on the markers selected in this paper. After confirmation of the identified compounds by acquisition of reference standards, the next step will be the development and validation of a target quantitative method, e.g. based on LC-MS/MS with triple quadrupole analyzer. This would offer a simple but efficient way to improve quality control in the olive oil industry using a widespread and less sophisticated MS/MS instrument, which is commonly available in quality control laboratories. Moreover, more information about the health benefits of detected compounds, as well

as their implications in the consumer decisions should be studied in a multidisciplinary work, including biochemists and/or medical researchers.

Acknowledgments

The authors acknowledge the financial support of Generalitat Valenciana, as research group of excellence (PROMETEO II/2014/ 023) and Collaborative Research on Environment and Food-Safety (ISIC/2012/016). The authors acknowledge InterCoop for Valencia Community EVOO samples and the information provided about the samples. This work has been developed with financial support from Fundacio Bancaixa (P1-1B2010-50 and P1-1B2013-70).

References

Ali, K., Maltese, F., Toepfer, R., Choi, Y. H., & Verpoorte, R. (2011). Metabolic characterization of Palatinate German white wines according to sensory attributes, varieties, and vintages using NMR spectroscopy and multivariate data analyses. *Journal of Biomolecular NMR*, 49(3-4), 255-266.

Angerosa, F., Servili, M., Selvaggini, R., Taticchi, A., Esposto, S., & Montedoro, G. (2004). Volatile compounds in virgin olive oil: Occurrence and their relationship with the quality. *Journal of Chromatography A*, 1054(1-2), 17-31.

Aparicio, R., & Aparicio-Ruız, R. (2000). Authentication of vegetable oils by chromatographic techniques. *Journal of Chromatography A*, 881(1e2), 93-104.

Aparicio, R., Morales, M. T., Aparicio-Ruiz, R., Tena, N., & García-Gonzalez, D. L. (2013). Authenticity of olive oil: Mapping and comparing official methods and promising alternatives. *Food Research International,* 54(2), 2025-2038.

Beltrán, M., Sánchez-Astudillo, M., Aparicio, R., & García-González, D. L. (2015). Geographical traceability of virgin olive oils from south-western Spain by their multi-elemental composition. *Food Chemistry*, 169, 350-357.

Camin, F., Pavone, A., Bontempo, L., Wehrens, R., Paolini, M., Faberi, A., et al. (2016). The use of IRMS, 1H NMR and chemical analysis to characterise Italian and imported Tunisian olive oils. *Food Chemistry*, 196, 98-105.

Cavalli, J.-F., Fernandez, X., Lizzani-Cuvelier, L., & Loiseau, A.-M. (2004). Characterization of volatile compounds of French and Spanish virgin olive oils by HSSPME: Identification of quality-freshness markers. *Food Chemistry*, 88(1), 151-157.

Cevallos-cevallos, J. M., Etxeberria, E., Danyluk, M. D., & Rodrick, G. E. (2009). Metabolomic analysis in food science: A review. *Trends in Food Science & Technology*, 20(11-12), 557-566.

Dais, P., & Hatzakis, E. (2013). Quality assessment and authentication of virgin olive oil by NMR spectroscopy: A critical review. *Analytica Chimica Acta*, 765, 1-27.

Díaz, R., Pozo, O. J., Sancho, J. V., & Hernández, F. (2014). Metabolomic approaches for orange origin discrimination by ultra-high performance liquid chromatography coupled to quadrupole time-of-flight mass spectrometry. *Food Chemistry*, 157, 84-93.

Do, T. K. T., Hadji-Minaglou, F., Antoniotti, S., & Fernandez, X. (2015). Authenticity of essential oils. *TrAC Trends in Analytical Chemistry*, 66, 146-157.

Erraach, Y., Sayadi, S., Gómez, A. C., & Parra-López, C. (2014). Consumer-stated preferences towards Protected Designation of Origin (PDO) labels in a traditional olive-oil-producing country: The case of Spain. *New Medit*, 13(4), 11-19.

Faria, M. A., Cunha, S. C., Paice, A. G., & Oliveira, M. B. P. P. (2010). Olives and olive oil in health and disease prevention. Olives and olive oil in health and disease prevention. Elsevier.

Fazio, C., & Ricciardiello, L. (2014). Components of the Mediterranean Diet with chemopreventive activity toward colorectal cancer. *Phytochemistry Reviews*, 13(4), 867-879.

Flath, R. A., Forrey, R. R., & Guadagni, D. G. (1973). Aroma components of olive oil. *Journal of Agricultural and Food Chemistry*, 21, 948-952.

Galeano Díaz, T., Durán Merás, I., Sánchez Casas, J., & Alexandre Franco, M. F. (2005). Characterization of virgin olive oils according to its triglycerides and sterols composition by chemometric methods. *Food Control*, 16(4), 339-347.

Gallart-Ayala, H., Chereau, S., Dervilly-Pinel, G., & Le Bizec, B. (2015). Potential of mass spectrometry metabolomics for chemical food safety. *Bioanalysis*, 7(1), 133-146.

Gamazo-Vázquez, J., García-Falcón, M. S., & Simal-Gándara, J. (2003). Control of contamination of olive oil by sunflower seed oil in bottling plants by GC-MS of fatty acid methyl esters. *Food Control*, 14(7), 463-467.

García-González, D. L., Luna, G., Morales, M. T., & Aparicio, R. (2009). Stepwise geographical traceability of virgin olive oils by chemical profiles using artificial neural network models. *European Journal of Lipid Science and Technology*, 111(10), 1003-1013.

Hu, W., Zhang, L., Li, P., Wang, X., Zhang, Q., Xu, B., et al. (2014). Characterization of volatile components in four vegetable oils by headspace two-dimensional comprehensive chromatography time-of-flight mass spectrometry. *Talanta*, 129, 629-635.

Jiménez, A., Aguilera, M. P., Beltrán, G., & Uceda, M. (2006). Application of solid-phase microextraction to virgin olive oil quality control. *Journal of Chromatography A,* 1121(1), 140-144.

Lísa, M., & Holcapek, M. (2008). Triacylglycerols profiling in plant oils important in food industry, dietetics and cosmetics using high-performance liquid chromatography-atmospheric pressure chemical ionization mass spectrometry. *Journal of Chromatography. A*, 1198-1199, 115-130.

Longobardi, F., Ventrella, A., Casiello, G., Sacco, D., Tasioula-Margari, M., Kiritsakis, A. K., et al. (2012). Characterisation of the geographical origin of Western Greek virgin olive oils based on instrumental and multivariate 358 statistical analysis. *Food Chemistry*, 133(1), 169-175.

Maggio, R. M., Cerretani, L., Chiavaro, E., Kaufman, T. S., & Bendini, A. (2010). A novel chemometric strategy for the estimation of extra virgin olive oil adulteration with edible oils. *Food Control*, 21(6), 890-895.

Peres, F., Jelen, H. H., Majcher, M. M., Arraias, M., Martins, L. L., & Ferreira-Dias, S. (2013). Characterization of aroma compounds in Portuguese extra virgin olive oils from Galega Vulgar and Cobrançosa cultivars using GC-O and GCxGC-ToFMS. *Food Research International,* 54(2), 1979-1986.

Portarena, S., Gavrichkova, O., Lauteri, M., & Brugnoli, E. (2014). Authentication and traceability of Italian extra-virgin olive oils by means of stable isotopes techniques. *Food Chemistry*, 164, 12-16.

Pouliarekou, E., Bakeda, A., Tasioula-Margari, M., Kontakos, S., longobardi, F., & Kontominas, M. G. (2011). Characterization and classification of Western Greek olive oils according to cultivar and geographical origin based on volatile compounds. Journal of Chromatography A, 1218, 7534-7542.

Purcaro, G., Cordero, C., Liberto, E., Bicchi, C., & Conte, L. S. (2014). Toward a definition of blueprint of virgin olive oil by comprehensive two-dimensional gas chromatography. *Journal of Chromatography. A,* 1334, 101-111.

Rojo, D., Canuto, G. A. B., Castilho-Martins, E. A., Tavares, M. F. M., Barbas, C., Lopez- Gonzálvez, A., et al. (2015). A multiplatform metabolomic approach to the basis of antimonial action and resistance in leishmania infantum. *PloS One*, 10(7), e0130675.

Smith, C. A., Want, E. J., O'Maille, G., Abagyan, R., & Siuzdak, G. (2006). XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Analytical Chemistry*, 78(3), 779-787.

Vaclavik, L., Cajka, T., Hrbek, V., & Hajslova, J. (2009). Ambient mass spectrometry employing direct analysis in real time (DART) ion source for olive oil quality and authenticity assessment. *Analytica Chimica Acta*, 645(1-2), 56-63.

Vasto, S., Buscemi, S., Barera, A., Di Carlo, M., Accardi, G., & Caruso, C. (2014). Mediterranean diet and healthy ageing: A sicilian perspective. *Gerontology*, 60(6), 508-518.

Vergara-Barberan, M., Lerma-García, M. J., Herrero-Martínez, J. M., & Simó-Alfonso, E. F. (2015). Cultivar discrimination of Spanish olives by using direct FTIR data combined with linear discriminant analysis. *European Journal of Lipid Science and Technology*, 117(9), 1473-1479.

Supplementary material

Instrumental conditions

The polar fraction was analyzed with mobile phase A= H_2O 0.01% HCOOH and B=MeOH 0.01% HCOOH. The percentage of organic modifier (B) was changed as follows: 0 min, 10%; 14 min, 90%; 20 min, 100%; 20.01 min, 10% in a total run time of 22 min for the positive ionization mode and 0 min, 10%; 5 min, 30%; 14 min, 60%; 14.50 min, 70%; 14.51 min, 100%; 16.00 min, 100%; 16.01 min, 10% in a total run time of 18 min for the negative ionization mode. The injection volume was 10 μL in both cases.

The non-polar fraction was analyzed using A= H_2O:ACN (15:85, v/v) 0.01% HCOOH 0.5 mM NH_4Ac and B= BuOH 0.01% HCOOH 0.5 mM NH_4Ac. The percentage of organic modifier (B) was changed as follows: 0 min, 0%; 3 min, 10%; 5 min, 50%; 6 min, 55%; 9 min, 60%; 11 min, 70% in a total run time of 13 min in both ionization modes. After each injection, a cleaning gradient was run to avoid memory effects due to relatively nonpolar molecules trapped in the injection valve. The percentage of organic modifier (B) changed as follows: 0 min, 50%; 4.50 min, 70%; 4.51 min, 0% in a total run time of 6 minutes. The injection volume for both positive and negative non-polar analyses was 5 μL.

Nitrogen was used as desolvation gas and nebulizing gas. The desolvation gas flow was set at 1000 L/h, and the cone gas was set at 80 L/h. The desolvation gas temperature was set to 600 ℃, the source temperature to 120 ℃ and the column temperature was set to 40 ℃ for polar fraction and 60 ℃ for non-polar one. A capillary voltage of 0.7 kV and 1.5 kV for positive and negative ion modes, respectively and a cone voltage of 25 V were used. MS data were acquired over an *m/z* range of 50-1200. Collision gas was argon 99.995% (Praxair, Valencia, Spain).

For MS^E experiments, two acquisition functions with different collision energies were created: the low energy (LE) function, selecting as collision energy 4eV, and the high energy (HE) function, with a collision energy ramp from 15 to 40 eV. The LE and HE functions settings were for both a scan time of 0.3 s and an inter-scan delay of 0.05 s.

Calibrations were conducted from *m/z* 50 to 1200 with a 1:1 mixture of 0.05 M NaOH:5% HCOOH diluted (1:25) with H_2O:ACN (20:80), at a flow rate of 10 µL/min. For automated accurate mass measurement, a Leucine-enkephalin solution (0.5 µg/mL) in ACN:H_2O (50:50) at 0.1% HCOOH was pumped at 30 µL/min through the lock-spray needle every 30 seconds, with a scan time of 0.3 seconds. The (de)protonated molecule of Leucine-enkephalin, at *m/z* 556.2771 in positive mode and *m/z* 554.2615 in negative mode, was used for recalibrating the mass axis during the injection and to ensure a robust accurate mass along the time.

Supplementary figures

Figure S1: Different regions of EVOO sampling. In red, *Cataluña* (10), in orange, *Bajo Aragón (5), in purple, Navarra-rioja (8), in brown, Toledo (5), in green Comunidad Valenciana (33) and in Blue Andalucia (29).*

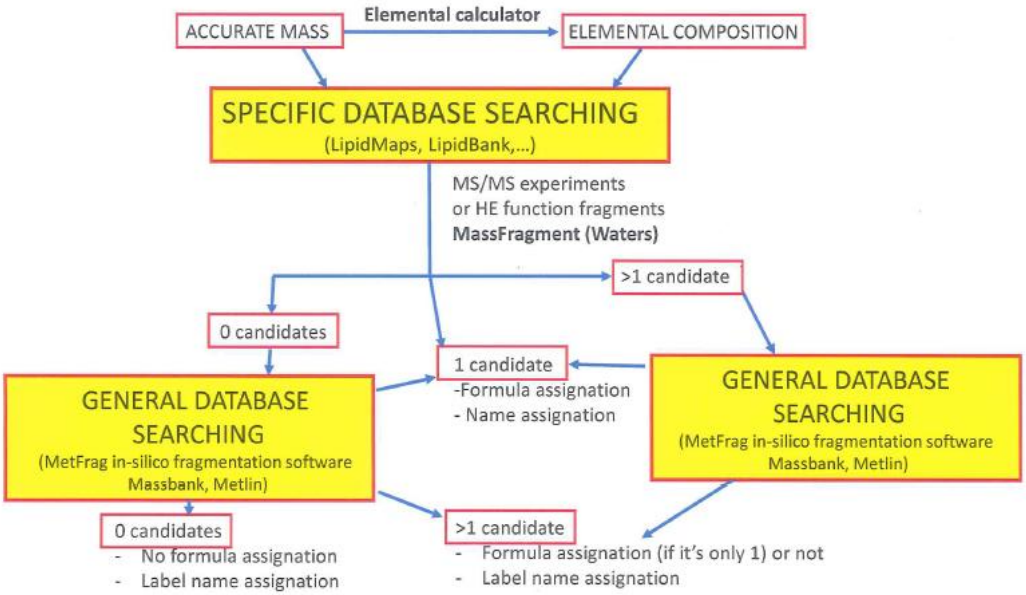
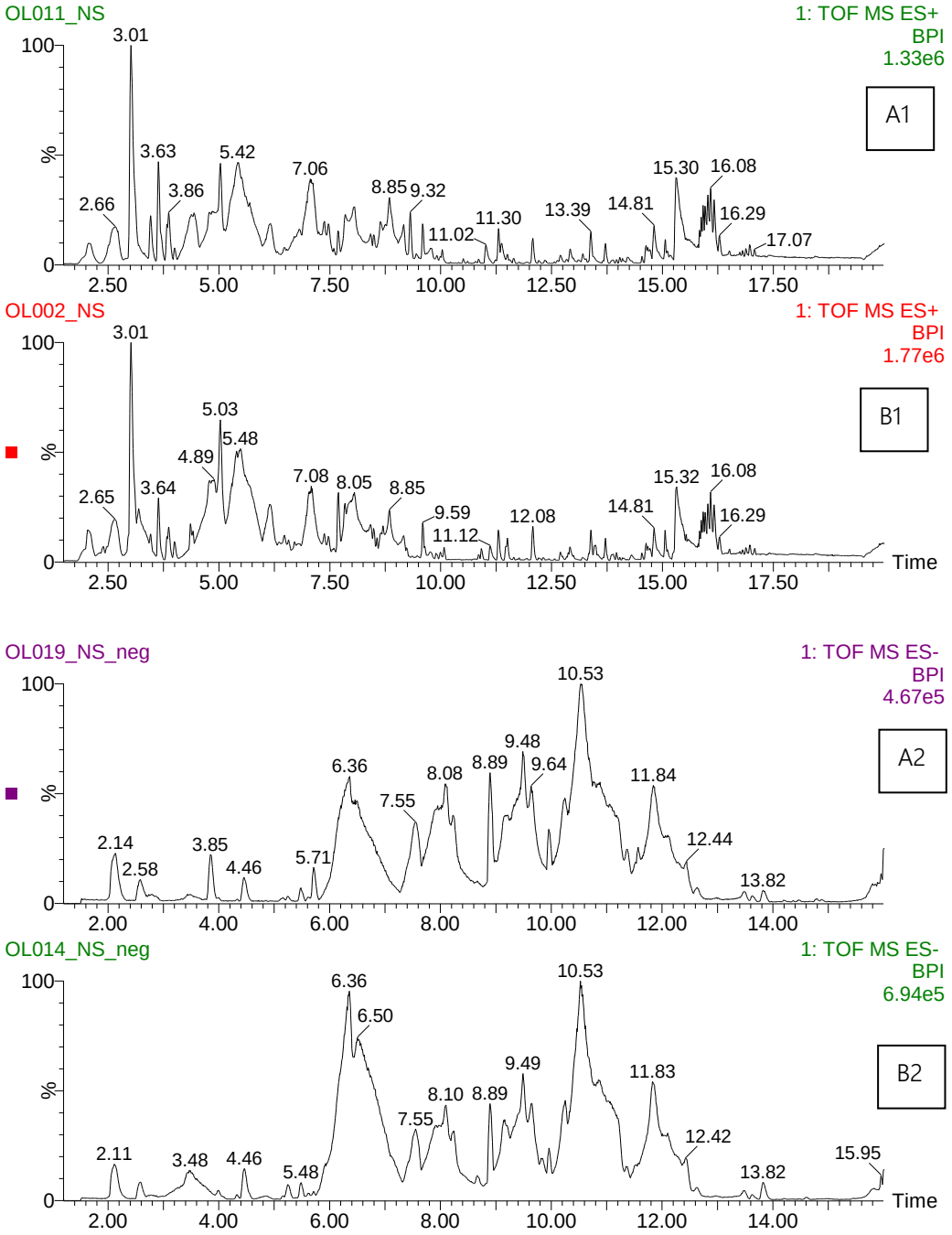


Figure S2: Elucidation workflow followed in this study.


OL011_NS
1: TOF MS ES+
BPI
1.33e6
A1
OL002_NS
1: TOF MS ES+
BPI
1.77e6
B1
OL019_NS_neg
1: TOF MS ES-
BPI
4.67e5
A2
OL014_NS_neg
1: TOF MS ES-
BPI
6.94e5
B2
Time

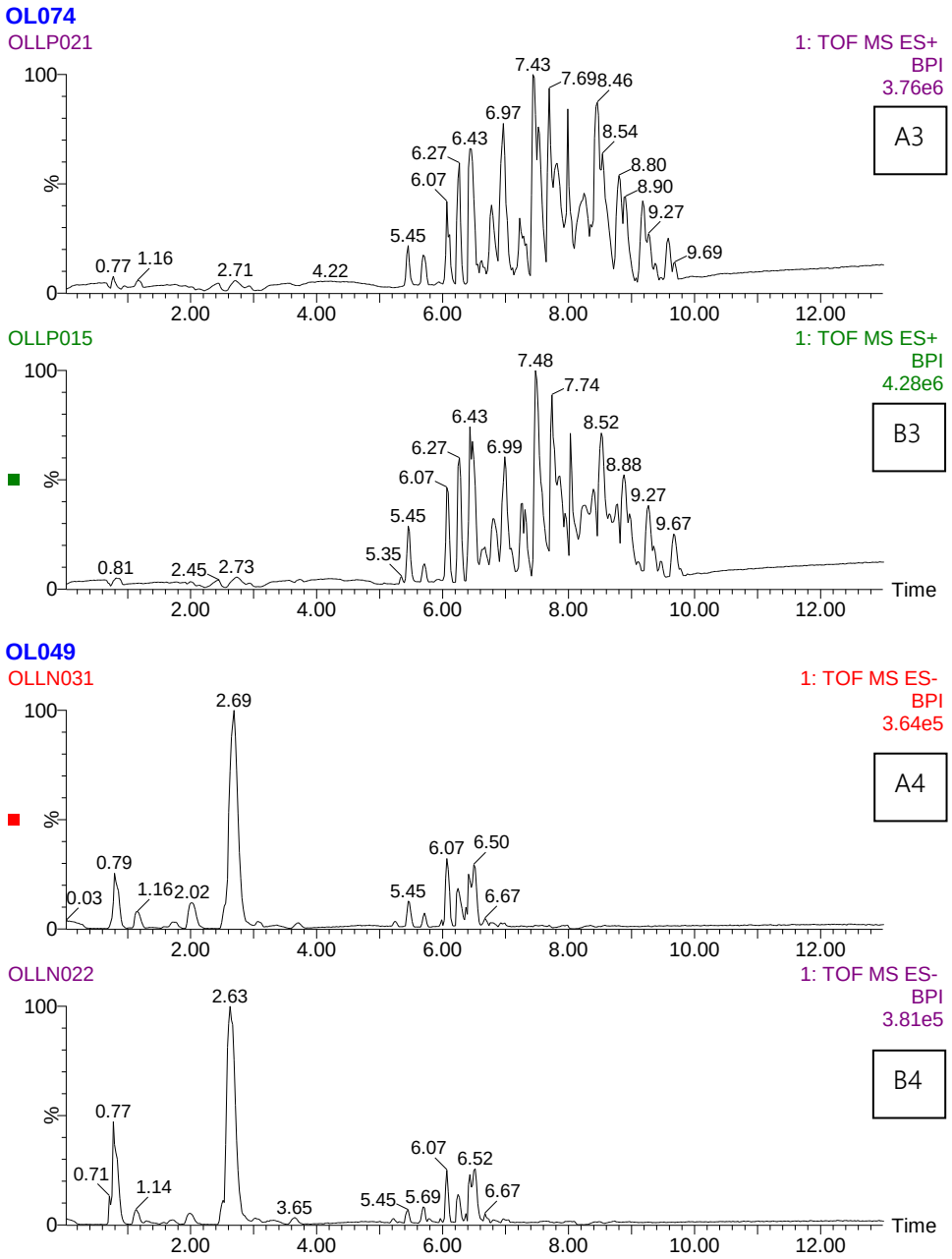


Figure S3: BPI chromatograms for two EVOOs collected in different Spanish regions (A, Aragón; B, Cataluña). 1: polar fraction in ESI positive, 2: polar fraction in ESI negative, 3: non polar fraction in ESI positive and 4: non polar fraction in ESI negative ionization modes

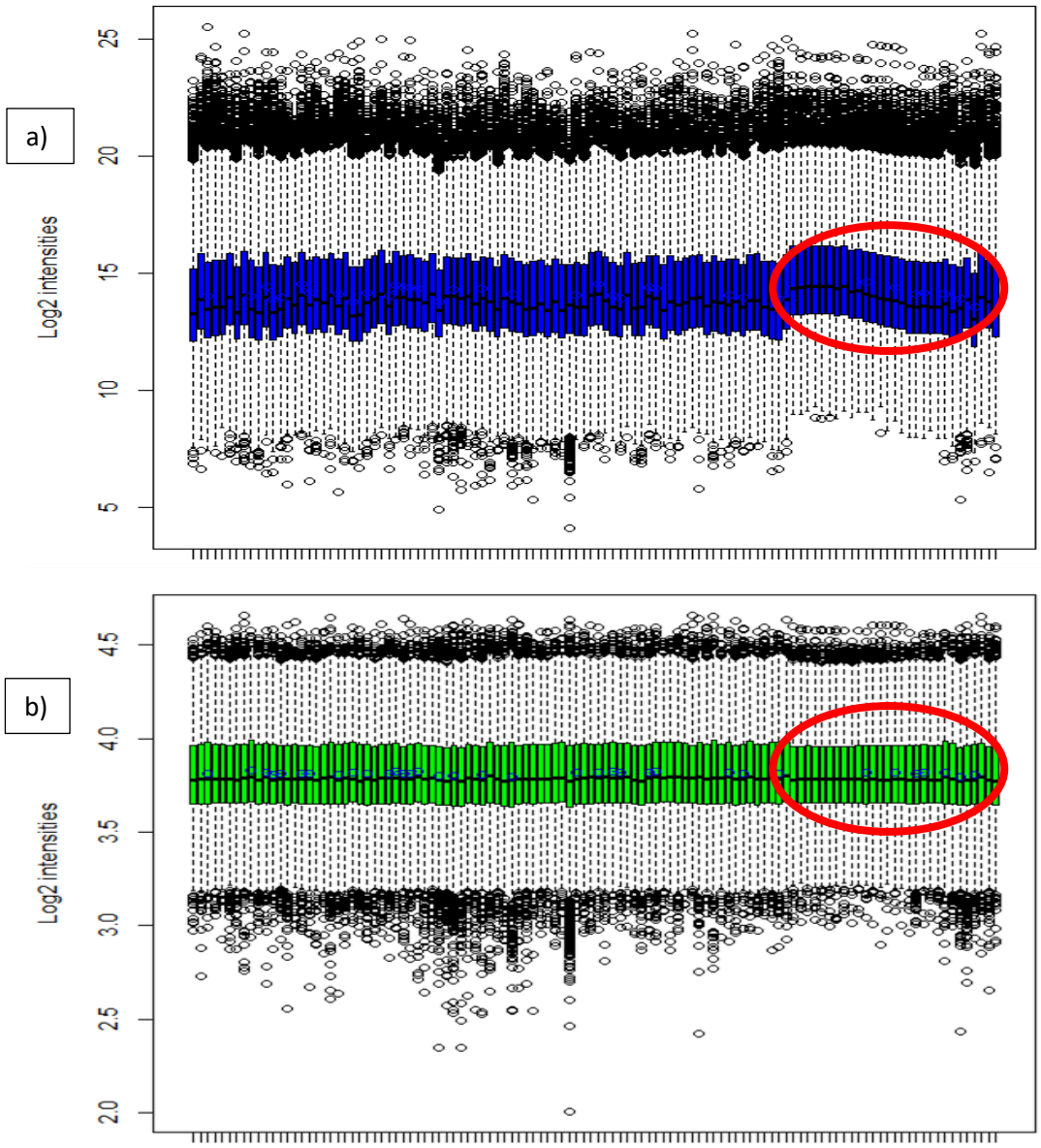
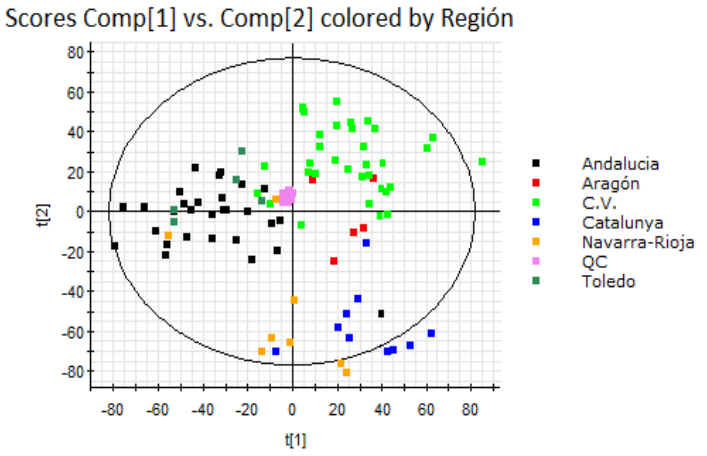
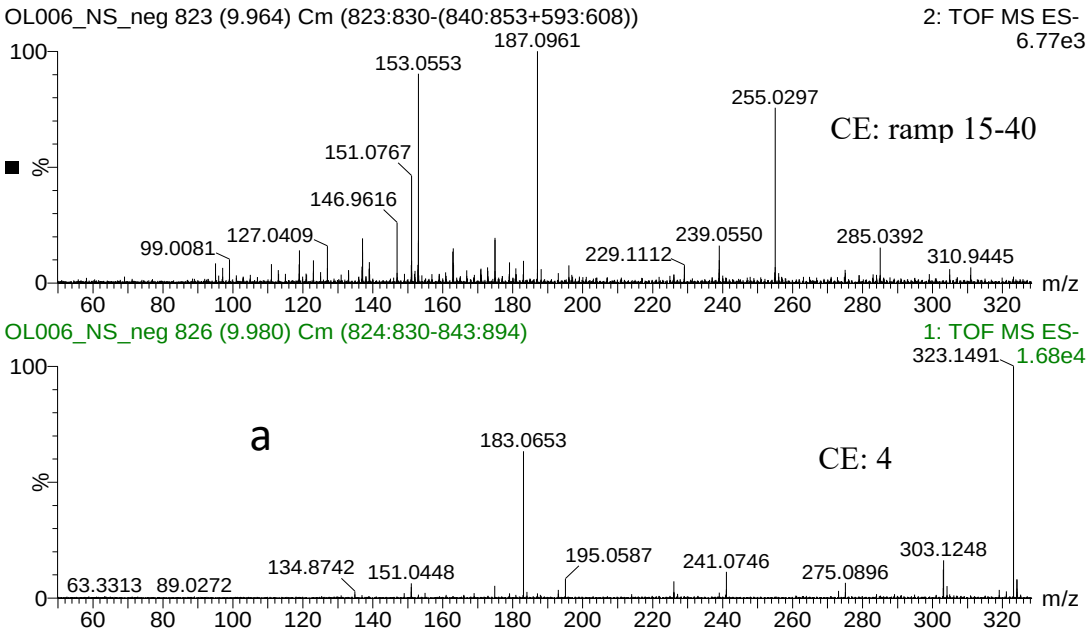


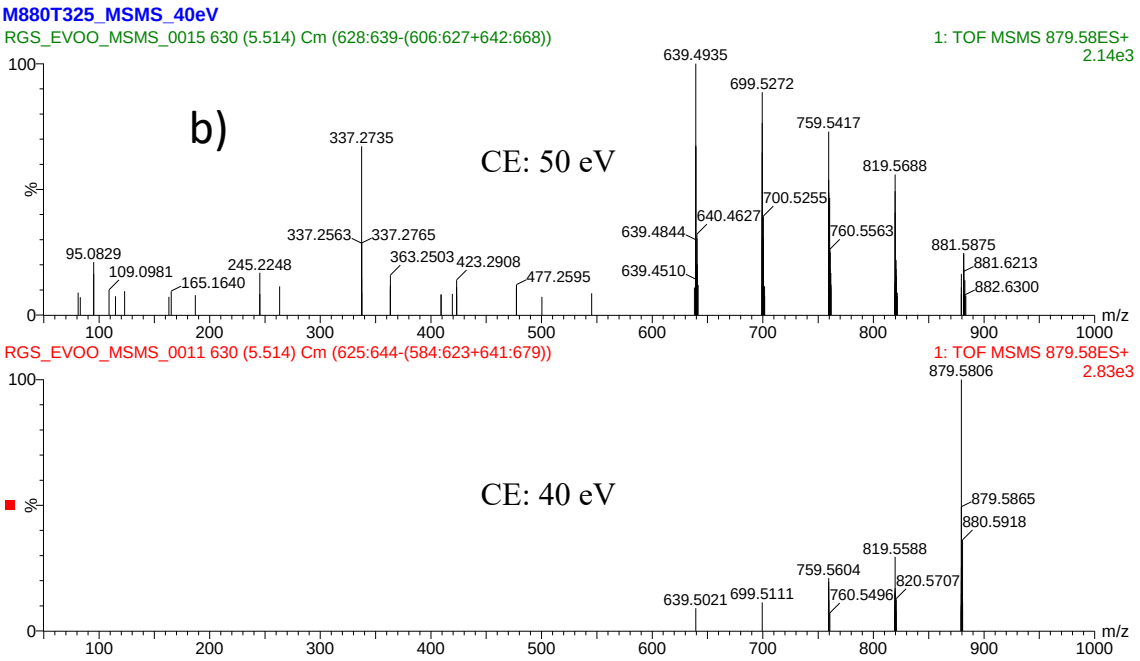
Figure S4: Box plots of the Log_2 intensities of the different samples in reverse phase data a) before and b) after normalization, in positive ionization mode. QC samples are highlighted.

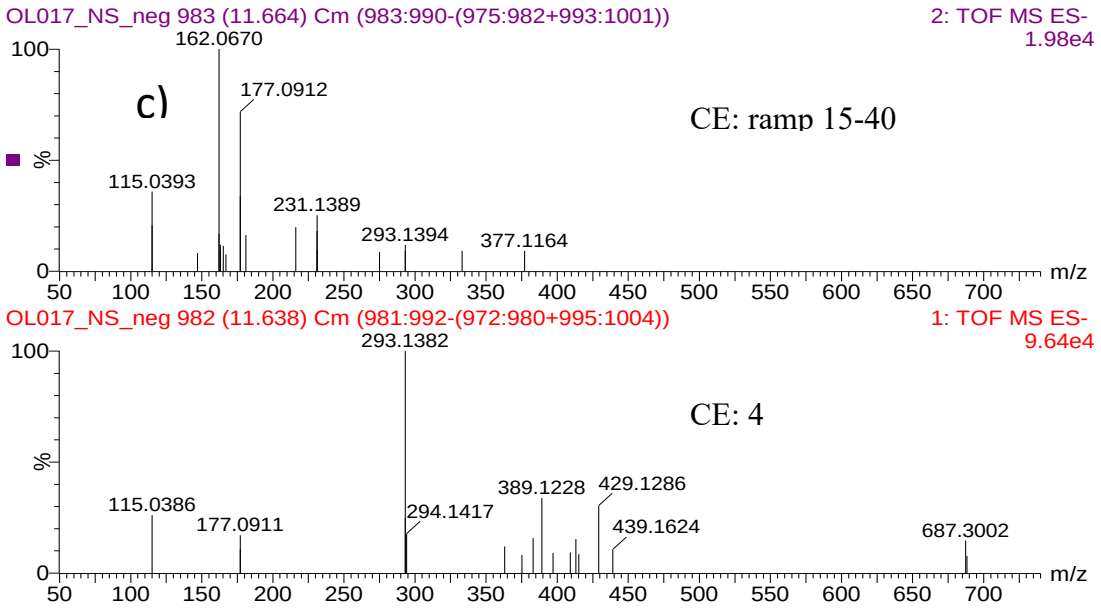


Comp No.	M2.R2Y(cum)	% Explained Variance
1	0.160236	16.02
2	0.309674	30.97
3	0.425975	42.60
4	0.524308	52.43
5	0.590249	61.89
6	0.698425	69.84
7	0.757756	75.77
8	0.810609	81.85
9	0.855178	85.51
10	0.879632	87.96
11	0.908964	90.90
12	0.931007	93.10

Figure S5: *Left*: Score Plot from PLS-DA. QC samples are highlighted in the center of the plot. *Right*: Total variance explained by the first 12 components in PLS-DA.




M880T325_MSMS_40eV
RGS_EVOO_MSMS_0015 630 (5.514) Cm (628:639-(606:627+642:668))
1: TOF MSMS 879.58ES+
2.14e3
b)
CE: 50 eV
639.4935
699.5272
759.5417
819.5688
337.2735
700.5255
640.4627
639.4844
760.5563
337.2563
337.2765
881.5875
881.6213
882.6300
363.2503
423.2908
95.0829
109.0981
245.2248
165.1640
639.4510
477.2595
RGS_EVOO_MSMS_0011 630 (5.514) Cm (625:644-(584:623+641:679))
1: TOF MSMS 879.58ES+
2.83e3
879.5806
CE: 40 eV
879.5865
880.5918
819.5588
759.5604
820.5707
639.5021
699.5111
760.5496
%
m/z


OL017_NS_neg 983 (11.664) Cm (983:990-(975:982+993:1001))
2: TOF MS ES-
1.98e4
c)
CE: ramp 15-40
162.0670
177.0912
115.0393
231.1389
293.1394
377.1164
OL017_NS_neg 982 (11.638) Cm (981:992-(972:980+995:1004))
1: TOF MS ES-
9.64e4
293.1382
CE: 4
389.1228
429.1286
115.0386
294.1417
177.0911
439.1624
687.3002
%
m/z

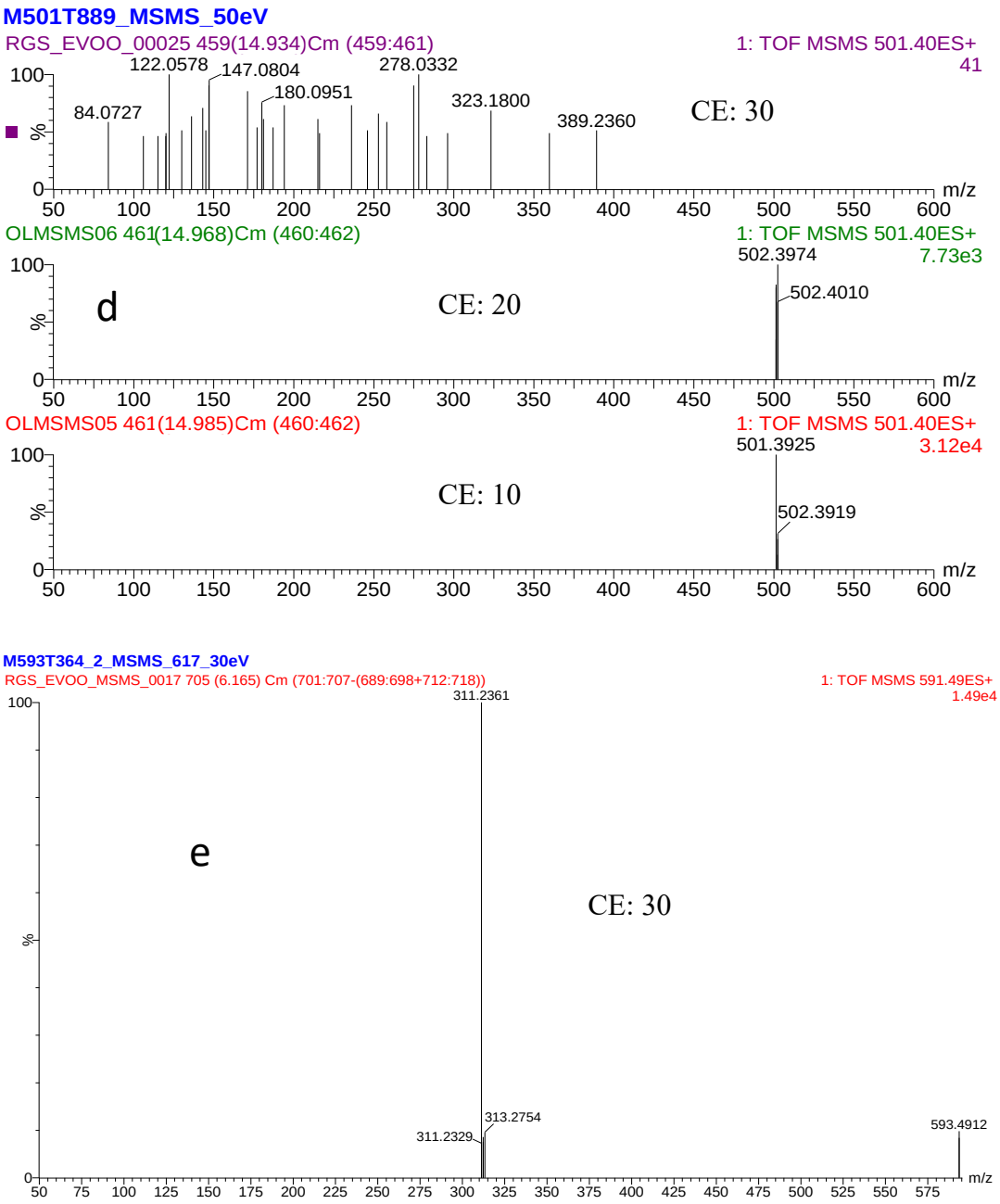


Figure S6: Product ion spectra or Low and High Energy spectra for unidentified compounds. (a) Compound 8, (b) Compound 9, (c) Compound 10, (d) Compound 11 and (e) Compound 12.

III.2: Artículo científico 4

FOOD ADDITIVES & CONTAMINANTS: PART A, 2018
VOL. 35, NO. 3, 395–403
https://doi.org/10.1080/19440049.2017.1416679

The classification of almonds (*Prunus dulcis*) by country and variety using UHPLC-HRMS-based untargeted metabolomics

R. Gil Solsona, C. Boix, M. Ibáñez and J. V. Sancho

Research Institute for Pesticides and Water (IUPA), University Jaume I, Castellón, Spain

ABSTRACT

The aim of this study was to use an untargeted UHPLC-HRMS-based metabolomics approach allowing discrimination between almonds based on their origin and variety. Samples were homogenised, extracted with ACN:H_2O (80:20) containing 0.1% HCOOH and injected in a UHPLC-QTOF instrument in both positive and negative ionisation modes. Principal component analysis (PCA) was performed to ensure the absence of outliers. Partial least squares – discriminant analysis (PLS-DA) was employed to create and validate the models for country (with five different compounds) and variety (with 20 features), showing more than 95% accuracy. Additional samples were injected and the model was evaluated with blind samples, with more than 95% of samples being correctly classified using both models. MS/MS experiments were carried out to tentatively elucidate the highlighted marker compounds (pyranosides, peptides or amino acids, among others). This study has shown the potential of high-resolution mass spectrometry to perform and validate classification models, also providing information concerning the identification of the unexpected biomarkers which showed the highest discriminant power.

ARTICLE HISTORY
Received 14 September 2017
Accepted 15 November 2017

KEYWORDS
Almond; untargeted metabolomics; UHPLC; high-resolution mass spectrometry; PLS-DA

Introduction

Nuts and olive oil are considered fundamental in the Mediterranean diet, not only due to their healthy lipid profile (Kodad and Socias i Company 2008), but also the presence of antioxidants, phenols, flavonoids (Alasalvar and Bolling 2015), vitamins and/or phytosterols (Bullo et al. 2011). Their consumption has been related to several benefits for human health, proved by published clinical diet trials such as those showing lower levels of blood cholesterol (Hyson et al. 2002) as well as reduced coronary heart disease (Ros 2010), reduction of serum uric acid (Jamshed et al. 2015) and also linked with prebiotic effects on gut microbiota (Ukhanova et al. 2014), which makes almonds or nuts in general an important part of a balanced diet.

The almond (*Prunus dulcis*) is a nut largely consumed in the Mediterranean diet (Bullo et al. 2011), considered one of the healthiest diets in the world. However, almonds have different fatty acid profiles depending on their origin and variety, which affects their stability against rancidity during storage or transport (Kodad and Socias i Company 2008), directly affecting their flavour and also influencing their price. Apart from this problem, two typical sweet almond-derived products (turron and marzipan) require a minimum amount of almond, protein and fat to be considered "high quality" products (Romero 2014). In this sense, genotype or growing region is really important for nutrient composition in almonds (Yada et al. 2011). Spanish almonds provide different characteristics and flavour to these products making them highly appreciated by the turron and marzipan industries. Thus, it becomes important to develop methods for origin and variety control to give producers analytical tools to ensure the authenticity of origin and/or variety of almonds. In this sense, different experiments have been performed to obtain fingerprints which allow the classification of almonds based on their geographical origin. Most of them are based on lipid profiling, the most studied compounds being fatty acids (Amorello et al. 2016), triacylglycerols and phospholipids (Shen et al. 2013; Petroselli et al. 2015), or also in combination with tocopherol analysis (Barreira et al. 2012). The main problem with these studies is the lack of validation of the

CONTACT J. V. Sancho sanchoj@uji.es Research Institute for Pesticides and Water (IUPA), University Jaume I, Castellón, Spain
Color versions of one or more of the figures in this article can be found online at www.tandfonline.com/TFAC
Supplemental data can be accessed here.

The classification of almonds (*Prunus Dulcis*) by country and variety using UHPLC-HRMS-based untargeted metabolomics.

R. Gil-Solsona, C. Boix, M. Ibáñez, J. V. Sancho*

Research Institute for Pesticides and Water (IUPA), University Jaume I, Castellón, Spain.

Abstract

The aim of this study was the use of an untargeted UHPLC-HRMS-based metabolomics approach which allows the discrimination between almonds based on their origin and variety. Samples were homogenized, extracted with ACN:H_2O (80:20), 0.1% HCOOH and injected in a UHPLC-QTOF instrument in both positive and negative ionization modes. PCA was performed to ensure the absence of outliers. PLS-DA was employed to create and validate the models for country (with 5 different compounds) and variety (with 20 features), showing more than 95% accuracy. Additional samples were injected and the model was evaluated with blind samples, observing more than 95% of correctly classified samples in both models. MS/MS experiments were carried out to tentatively elucidate the highlighted markers (piranosides, peptides or aminoacids amongst others). This study has shown the potential of HRMS to perform and validate classification models, also providing information for the identification of the unexpected biomarkers, which showed the highest discriminant power.

Introduction

Nuts and olive oil are considered fundamental in the Mediterranean diet, not only due to their healthy lipid profile (*Kodad & Socias i Company, 2008*), but also for the presence of antioxidants, phenols, flavonoids (*Alasalvar & Bolling, 2015*), vitamins and/or phytosterols (*Bullo, Lamuela-Raventos, & Salas-Salvado, 2011)*. Their consumption has been related to several benefits for human health, proved by clinical diet trials in the literature such as lower levels of blood cholesterol (*Hyson, Schneeman, & Davis, 2002*) as well as reduced coronary heart disease (*Ros & Emilio, 2010*), reduction of serum uric acid (*Jamshed et al., 2015*) and also linked with prebiotic effects on gut microbiota *(Ukhanova et al., 2014*) which makes almonds, or nuts in general, an important part of a balanced diet.

Almonds (*Prunus Dulcis*) is a nut largely consumed in the Mediterranean diet (*Bullo et al., 2011*), considered one of the healthiest diets in the world. However, almonds have different fatty acid profiles depending on their origin and variety, which alters their stability against rancidification during storage or transport (*Kodad & Socias i Company, 2008*), directly affecting their flavour and also affecting their prize. Apart from this problem, two typical sweet almond derived products (turron and marzipan) requires a minimum amount of almond, protein and fat to be considered "high quality" products (*Romero, 2014*). In this sense, genotype or growing region is really important for nutrient composition in almonds (*Yada, Lapsley, & Huang, 2011*). Spanish almonds provides different characteristics and flavour to these products making them highly appreciated in turron and marzipan industries. Thus, it becomes important to develop methods for origin and variety control to give producers analytical tools to ensure almonds origin and/or variety. In this sense, different experiments have been performed to obtain fingerprints which allows the classification of almonds based on their geographical origin. Most of them are based on lipid profiling, being the most studied compounds fatty acids (*Amorello, Orecchio, Pace, & Barreca, 2016*), triacylglycerols and phospholipids (*Petroselli et al., 2015; Shen et al., 2013*), or also in combination with tocopherol analysis (*Barreira et al., 2012*). The main problem of these studies was the lack of validation of the developed methodology and the fact that the selected compounds could not be the best markers to distinguish between them as a targeted approach was used. Untargeted metabolomics could be a good option to highlight the best compounds to discriminate samples in different scenarios like animal diets (*Ruiz-Aracama, Lommen, Huber, Van De Vijver, & Hoogenboom, 2011*) or food traceability, as almond classification by cultivar (*Beltrán Sanahuja, Ramos Santonja, Grané Teruel, Martín Carratalá, & Garrigós Selva, 2011*). Their main drawback were, as previously commented for targeted approaches, the lack of extra validation steps to confirm that their promising markers are robust along the time.

The untargeted metabolomics approach becomes very useful for food control (*Gil-Solsona et al., 2016; Sales et al., 2017*). In this sense, the powerful chromatographic techniques coupled to high resolution MS (HRMS) (*Emwas, 2015*) provide the perfect tool for this aim, as it has been demonstrated by their increasing use in the last years in food authenticity and control (*Castro-Puyana & Herrero, 2013; Rubert, Zachariasova, & Hajslova, 2015*).

The main aim of this research was to investigate the applicability of untargeted metabolomics approaches using ultra high performance liquid chromatography (UHPLC) coupled to HRMS to classify almond samples according to their country of origin as well as variety. For this purpose, almonds from Spain and USA were employed. Sample extracts were injected and after multivariate analysis the most relevant compounds were highlighted with Variable Importance in Projection (VIP) selection method. Partial Least Square-Discriminant Analysis (PLS-DA) was employed to create both models, which were validated with samples from a second season employed as a system challenge (*Riedl, Esslinger, & Fauhl-Hassek, 2015*). MS/MS experiments were performed for highlighted markers and tentatively elucidated with the help of online databases.

Materials a& methods

Reagents and chemicals

HPLC-grade water was obtained from a Mili-Q water purification system (Millipore Ltd., Bedford, MA, USA). HPLC-grade methanol (MeOH), HPLC-supergradient ACN, sodium hydroxide (>99%) and ammonium acetate (NH_4Ac) reagent grade were obtained from Scharlab (Barcelona, Spain). Leucine-enkephalin (mass-axis calibration) and formic acid (mobile phase modifier) were purchased from Sigma-Aldrich.

Sampling

Spanish almond of different varieties (*Bitter almond, Belona, Carrerona, Comuna, Ferranduel, Guara, Largueta and Marcona*) were purchased from Frusema company (*Albocasser, Castellón, Spain*). Almonds from USA (*Bute-padre, California and Non-Pareil*) were obtained from FruSecs company (*Albocasser, Castellón, Spain*). In a second season, samples were also obtained from *Frusema* and *FruSecs*. In this case, an additional Spanish variety, *Soleta*, was also sampled. A total of 62 sample packages containing 100 g of an individual variety were employed.

Sample processing

Raw samples (100 g) were triturated and homogenized. 2.5 g of sample were weighted and mixed with 10 mL of ACN:H_2O (80:20) 0.1% HCOOH. After mechanically shaking for 90 min, extracts were sonicated for 15 minutes and centrifuged for 10 min at 4.500 g. The supernatant was 4-fold diluted with Milli-Q water and stored at -24 ºC until their analysis. A pool of all the extracts was also performed, named QC, to obtain an average extract of our sample set. This pool was used for column stabilization (by injecting 10 QC samples at the beginning of each sample batch), and to control possible instrumental signal variation along the sequence.

UHPLC-HRMS

A Waters Acquity UPLC system (*Waters, Milford, MA, USA*) was coupled to a hybrid quadrupole-TOF mass spectrometer (Xevo G2 QTOF, *Waters, Manchester, UK*), using a Z-spray-ESI interface operating in both positive and negative ionization modes. The UHPLC separation was performed using a CORTECS® C18 fused-core 2.7 µm particle size analytical column 100 x 2.1 mm (Waters) at 300 µL/min flow rate. The separation was performed using H2O 0.01% HCOOH as weak mobile phase (A) and MeOH 0.01 % HCOOH as strong mobile phase (B). The percentage of B was changed from 10 % at 0 min, to 90 % at 14 min, 90 % at 16 min and 10 % at 16.01 min, with a total run time of 18 min. Injection volume was 10 µL. Nitrogen was used as both the desolvation gas and the nebulizing gas. A capillary voltage of 0.7 kV and 1.5 kV for positive and negative ion modes, respectively, and cone voltage of 25 V were used. MS data were acquired over a *m/z* range of 50-1200. TOF-MS resolution was approximately 20000 at full width half maximum at *m/z* 556.2771. Collision gas was argon 99.995% (Praxair, Valencia, Spain). The desolvation gas flow was set at 1000 L/h, and the cone gas was set at 80 L/h. The desolvation gas temperature was set to 600 ℃, the source temperature to 130 ℃ and the column temperature was set to 40 ℃.

For MSE experiments, two acquisition functions with different collision energies were created. The low energy (LE) function, with a fixed collision energy of 4 eV, and the high energy (HE) function, with a collision energy ramp ranging from 15 to 40 eV in order to obtain the (de)protonated ion from LE function and a wide range of fragment ions from the HE function. Both LE and HE

functions used a scan time of 0.3 s with an inter-scan delay of 0.05 s and were applied in the same injection simultaneously.

MS/MS experiments were carried out in the same conditions with different collision energies depending on the fragmentation observed for each compound. Calibrations were conducted from *m/z* 50 to 1200 with a 1:1 mixture of 0.05 M NaOH:5 % HCOOH diluted (1:25) with H2O:ACN (20:80), at a flow rate of 10 µL/min. For automated accurate mass measurement, a leucine-enkephalin solution (2 µg/mL) in ACN:H_2O (50:50) at 0.1% HCOOH was pumped at 20 µL/min through the lock-spray needle and measured every 30 s, with a scan time of 0.3 s. The (de)protonated molecule of leucine-enkephalin, at *m/z* 556.2771 in positive mode and *m/z* 554.2615 in negative mode was used for recalibrating the mass axis during the injection and to ensure a robust accurate mass along time.

Data processing

The untargeted metabolomics data workflow (**Figure S1**) starts converting LC-MS raw data from proprietary (*.raw, Waters Corp*.) to generic (*.cdf, NetCDF)* format using Databridge application (within MassLynx v 4.1; *Waters Corporation*) and processed using XCMS R package (https://xcmsonline.scripps.edu/) (*Smith, Want, O'Maille, Abagyan, & Siuzdak, 2006*). Centwave feature detection algorithm was employed for peak picking (peak width from 5 to 20 s, S/N ratio higher than 10 and mass tolerance of 15 ppm) to convert chromatograms into a list of detected features. It was followed by retention time alignment, to identify the same ion across different samples with slightly different retention time (around 10 seconds of difference). The aligned features were labeled as MxxxTyyy, where xxx corresponds to the nominal mass of the compound and yyy to the retention time in seconds. Mean centering was applied to normalize each data set, minimizing instrumental drifts between samples. Finally, log2 transformation was applied to the area of each detected signal to avoid heteroscedasticity, followed by Pareto scaling, which provides to the features their "statistical weight" regarding differences between groups and not depending on their total area.

Multivariate analysis

Principal component analysis (PCA, **Figure S2**) and PLS-DA were performed by means of the EZ-Info software (*Umetrics, Sweden*). Firstly, PCA was used to ensure the absence of outliers and the correct grouping of QC samples after normalization. PLS-DA was then applied to reduce dimensions in the dataset. By means of VIP filtering, the minimum required ions to achieve a good classification model were obtained.

The 75-80% of the dataset was employed to create the model, with samples from both first and second season, ensuring that the selected compounds were independent from the harvest year, while the other 20-25% were not included in the model creation. With these two groups, the model was validated in two steps. Firstly, a cross-validation was applied to control the model goodness and also an additional validation step was carried out with the 20-25% of the samples not included in the model. Statistical model gives two columns, the first one (Likely Classification) where the model assigns the sample unequivocally to the group obtained, and the second (Less Likely Classification) where model can provide no result, only one result and more than one result. Samples with only one result in this column are given as correct by us while the rest are treated as unknown.

Marker identification

The MS/MS spectra of the most significant metabolites at 10, 20, 30 and 40 eV were acquired and searched in online databases as METLIN (https://metlin.scripps.edu/landing-page.php?pgcontent=mainPage) or were in-silico tentatively elucidated with MetFrag Software (https://msbi.ipb-halle.de/MetFragBeta/), employing ChemSpider as chemical structure library When no hits were obtained, they were tried to be manually elucidated.

Results and discussion

Sample treatment

Almonds contains, regarding the polarity of the compounds, two fractions, a polar fraction (studied in this paper) and the less-polar fraction (mainly composed by lipids). The polar fraction requires polar solvents (water, methanol, acetonitrile…) while in order to extract less-polar compounds other kind of solvents should be employed (dichloromethane/methanol mixtures) as discussed in the literature (*Cevallos-cevallos, Etxeberria, Danyluk, & Rodrick, 2009*) or even butanol and 2-propanol when dealing directly with oils (*Gil-Solsona et al., 2016*).

For this reason, the non-polar compounds were extracted with ACN:H_2O (80:20) 0.1% HCOOH, proved as a good extraction solvent for a wide range of food matrices (*Beltrán et al., 2013*).

Data treatment

Both datasets (from positive and negative ionization modes) were joined in a single file, with a total of 1555 different ions. Data were then analyzed with PCA. At this point, QC samples are employed as an external standard to control the correct normalization. QC samples, as explained in Sample treatment section, are a pool of all the samples employed to perform the model. These samples, which have an average composition, should appear after normalization in the center of the PCA (non supervised method) and grouped, meaning that normalization steps (men centering, log2 and Pareto scaling) has corrected possible instrumental drifts and differences along the batch. In this case, after observing this correct QC grouping and the absence of outliers (**Figure S1**), PLS-DA models were created for country and variety classification.

It is important to observe that both datasets were joined in a single file in order to extract the best ions for the discrimination despite ionization mode. As we are employing an untargeted strategy, if one of both ionization modes better explains the differences between our groups, these ions will be preferably selected by VIP filtering. If not, we will ensure that the selected markers are the group that better explains the differences independently their ionization.

Country classification by PLS-DA

Initially, a model to differentiate the origin of the almonds was created. Samples were divided in three different groups, Spanish almonds (*Belona, Carrerona, Comuna, Ferranduel, Guara, Largueta, Marcona and Soleta*), USA almonds (*Bute-padre, California and Non-pareil*) and *Bitter* almonds, which showed a different behavior (see **Figure 1**). As it has been previously explained, both sample sets (first and second season) were mixed and 80% of the samples were employed to create the model, while the 20% were used for model validation.

The most important ions for the model were initially reduced down to 20 by means of their VIP value, ensuring that all the samples in the cross-validation were correctly classified. However, despite a soft ionization source was employed, more than one ion could be obtained for each marker compound. Therefore adducts and/or in-source fragments corresponding to the same marker were excluded based on mass accuracy as well as chromatographic profile. Finally, the total amount of ions was heavily reduced to just only 5 compounds/ions.

The PLS-DA model was created with these 5 ions with goodness-of-fit (R_2Y=0.848) and goodness-of-prediction (Q_2Y=0.771) for two first components. Then, it was validated in two steps, as recommended in the literature (Riedl et al., 2015). The first step was the cross-validation of the sample set employed to perform the model. Here, all the 50 samples, which were a mix of both seasons, were correctly labeled, ensuring the model robustness.

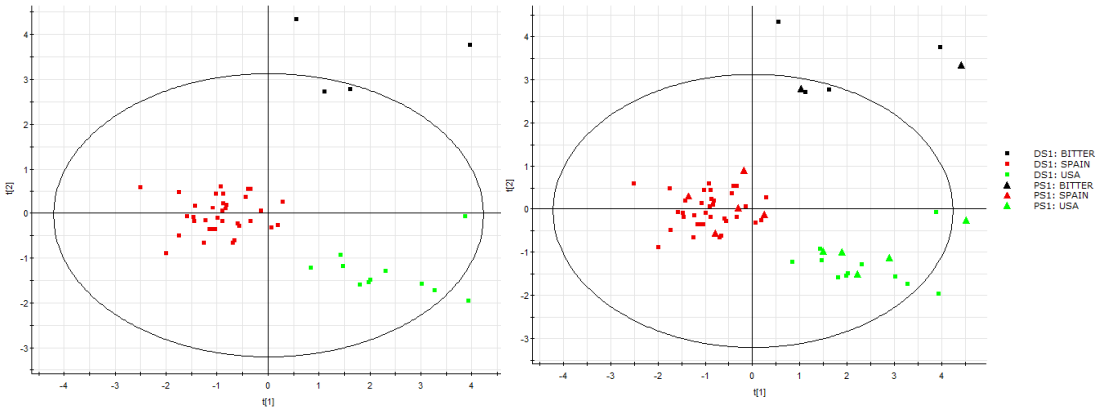


Figure 1: PLS-DA model for country classification. (A) Score plot of the first 2 components and (B) score plot with test samples.

To finally validate the model, 12 samples not included initially in the model creation (2 bitter almonds, 5 Spanish almonds and 5 USA almonds) were employed to test the model. All the samples were properly classified showing the successful applicability of the model.

Markers identity (see **Table 1**) was tentatively performed after MS/MS experiments. The most important product ions can be observed in the **Table S1**. Regarding the marker labeled as M318T239, after searching its accurate mass (*m/z* 318.2022) in *METLIN* database several hits were retrieved. However, the comparison of the empirical MS/MS spectrum with available spectra allowed us to tentatively elucidate it as the tripeptide (Val-Thr-Val). In a similar way, marker M298T178 was tentatively elucidated as (5'-deoxy-5'-(methylthio)adenosine).

Table 1: Markers selected for country discrimination.

Feature	Ionization mode	Ion	Tentative elucidation	Molecular formula	Exact mass / Accurate mass (error) $[M+H]^+$ or $[M-H]^-$	Retention time (min)
M448T119	POSITIVE	$[M+H]^+$	Diglucopyranosyl niacin	$C_{18}H_{25}NO_{12}$	448.1455 / 448.1458 (+0.3 mDa)	1.99
M298T178	POSITIVE	$[M+H]^+$	5'-Deoxy-5'-(methylthio)adenosine	$C_{11}H_{15}N_5O_3S$	298.0974 / 298.0959 (-1.5 mDa)	2.98
M293T201	NEGATIVE	$[M-H]^-$	Glucopyranosyl hydroxy caproic acid	$C_{12}H_{22}O_8$	293.1236 / 293.1231 (-0.5 mDa)	3.35
M318T239	POSITIVE	$[M+H]^+$	Val-Thr-Val	$C_{14}H_{27}N_3O_5$	318.2029 / 318.2022 (-0.7mDa)	3.94
M933T337	POSITIVE	$[M+NH_4]^+$	Amigdalin Hexose de-Hypoxanthine futalosine	$C_{40}H_{53}NO_{23}$	916.3087 / 916.3058 (-2.8 mDa)	4.98

M293T201 and M448T119 were tentatively identified as Glucopyrasonyl hydroxy caproic acid and Diglucopyranosyl niacin after observing in their MS/MS spectra losses corresponding to hexose groups ($C_6H_{10}O_5$). The remaining products ions allowed us to identify the corresponding aglycone moiety using METLIN database. Compound M933T337 was also searched in the METLIN database. But as no results were obtained, it was further evaluated using MetFrag In-silico fragmentation web tool searching for possible structures in Chemspider. Observing its fragmentation pattern, the consecutive neutral losses of amygdalin and hexose rendering a final product ion at *m/z* 297.0948, this marker was tentatively elucidated as de-Hypoxanthine futalosine conjugated with amygdalin and one hexose.

Spanish varieties classification by PLS-DA

In a second step, a classification model was created in order to differentiate among the Spanish varieties included in the first model. Samples were obtained after mixing almonds of different Spanish regions, always ensuring the same variety. This fact guarantees that the highlighted markers are robust for any Spanish almond, independently from their cultivar.

For this variety classification model, Spanish almonds were divided in seven different groups (*Belona, Carrerona, Comuna, Ferranduel, Guara, Largueta and Marcona*), employing the 75% (30 samples) of all the Spanish almonds to train the model and the remaining 25% (10 samples) to validate it, with at least one sample per variety. In the case of *Soleta* variety, as only one sample was obtained, it was only employed to test the model, evaluating potential misclassifications.

VIP filtering was applied to the whole table and features were checked to include only one ion per compound, typically the (de)protonated molecule or an adduct. Only 20 ions were necessary to build a model (see **Table 2**) with satisfactory goodness-of-fit (R_2Y=0.866) and goodness-of-prediction (Q_2Y=0.760) using eight components (**Figure 2A**).

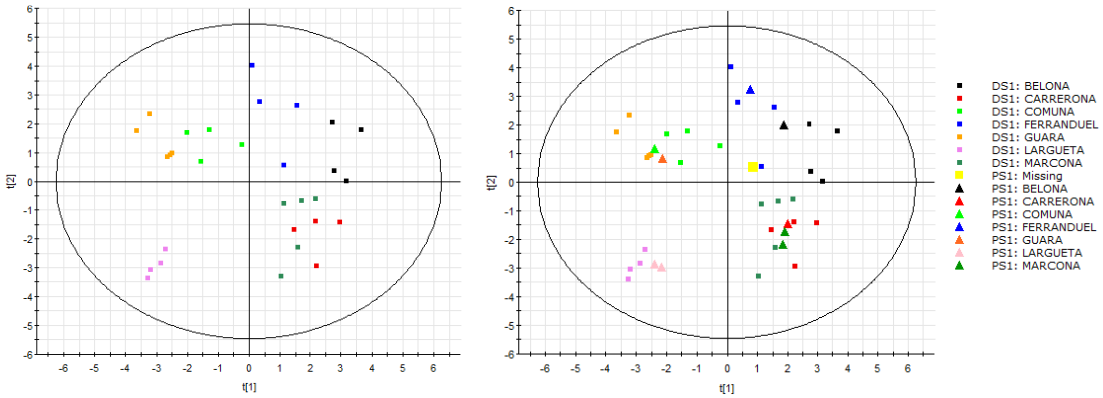


Figure 2: PLS-DA model for variety classification. (A) score plot of the first 2 components and (B) score plot with test samples. Soleta sample is labeled as unknown.

This classification model was again validated in two steps, a cross-validation, where 29 samples were correctly labeled, remaining only one as unknown. The final validation of the model was made with 10 samples not included in the initial model (*1 Belona, 1 Carrerona, 1 Comuna, 1 Ferranduel, 1 Guara. 2 Largueta, 2 Marcona and 1 Soleta). Soleta* was employed to test that samples

Table 2: Markers selected for variety discrimination.

Feature	Ionization mode	Ion	Tentative elucidation	Molecular formula	Exact mass / Accurate mass (error) $[M+H]^+$ or $[M-H]^-$	Retention time (min)
M503T55	NEGATIVE	$[M-H]^-$	Gentianose	$C_{18}H_{32}O_{16}$	503.1616 / 503.1612 (+0.4 mDa)	1.00
M377T68	NEGATIVE	$[M+Cl]^-$	Inulobiose	$C_{12}H_{22}O_{11}$	341.1070 / 341.1084 (-1.4 mDa)	1.23
M268T85	POSITIVE	$[M+H]^+$	Adenosine	$C_{10}H_{13}N_5O_4$	268.1045 / 268.1056 (-1.1 mDa)	1.46
M166T99	POSITIVE	$[M+H]^+$	L-Phenylalanine	$C_9H_{11}NO_2$	166.0867 / 166.0868 (-0.1 mDa)	1.64
M476T114	POSITIVE	$[M+H]^+$	Dihexosyl 2-ethylisonicotinic acid	$C_{20}H_{29}NO_{12}$	476.1786 / 476. 1768 (+1.8 mDa)	1.79
M577T127	NEGATIVE	$[M-H]^-$	Procyanidin B5	$C_{30}H_{26}O_{12}$	577.1334 / 577.1346 (-1.2 mDa)	2.05
M450T190	POSITIVE	$[M+NH_4]^+$	Benzyl gentobioside	$C_{19}H_{28}O_{11}$	433.1737 / 433.1710 (+2.7 mDa)	3.14
M313T217	POSITIVE	$[M+NH_4]^+$	Unknown	$C_{14}H_{17}NO_6$	296.1129 / 296.1107 (+2.2 mDa)	3.58
M289T214	NEGATIVE	$[M-H]^-$	Unknown	$C_{15}H_{14}O_6$	289.0706 / 289.0712 (-0.6 mDa)	3.6
M318T239	POSITIVE	$[M+H]^+$	Val-Thr-Val	$C_{14}H_{27}N_3O_5$	318.2029/ 318.2022 (-0.7 mDa)	3.87
M464T246	POSITIVE	$[M+NH_4]^+$	Hexosyl-2-phenylethyl glucopyranoside	$C_{20}H_{30}O_{11}$	447.1857 / 447.1866 (-0.9 mDa)	4.09
M548T340	POSITIVE	$[M+NH_4]^+$	Di Hexosyl -3-dimethylallyl-4-hydroxybenzoate	$C_{24}H_{34}O_{13}$	531.2090 / 531.2078 (+1.2 mDa)	5.65
M563T342	POSITIVE	$[M+H]^+$	Unknown	$C_{26}H_{30}N_2O_{12}$	563.1882 / 563.1877 (+0.5 mDa)	5.7
M540T362	POSITIVE	$[M+NH_4]^+$	Unknown	$C_{26}H_{34}O_{11}$	523.2181 / 523.2179 (+0.2 mDa)	6.05
M521T362	NEGATIVE	$[M-H]^-$	Unknown	$C_{26}H_{34}O_{11}$	521.2013 / 521.1983 (+3.0 mDa)	6.05
M287T340	NEGATIVE	$[M-H]^-$	7-hydroxy-3-(3,4,5-trihydroxyphenyl)-2,3-dihydro-4H-chromen-4-one	$C_{15}H_{12}O_6$	287.0542 / 287.0556 (-1.4 mDa)	6.07
M593T393	NEGATIVE	$[M-H]^-$	4,7-Dihydroxy-2-(4-hydroxyphenyl)-5-oxo-5H-chomen-3-yl-6-O-(6-deoxy-mannopyranosyl)-glucopyranoside	$C_{27}H_{30}O_{15}$	593.1505 / 593.1506 (-0.1 mDa)	6.56
M477T395	NEGATIVE	$[M-H]^-$	Unknown	$C_{22}H_{22}O_{12}$	477.1026 / 477.1033 (-0.7 mDa)	6.61
M647T403	POSITIVE	$[M+Na]^+$	Deoxyhexosyl Hexosyl Quercetin 3-methyl ester	$C_{28}H_{32}O_{16}$	625.1769 / 625.1769 (0.0 mDa)	6.72
M348T852	POSITIVE	$[M+H]^+$	Hexadecyl Methyl glycerol	$C_{20}H_{42}O_3$	331.3202 / 331.3212 (-1.0 mDa)	14.06

from varieties not included in the model were not wrongly labeled. 8 samples were correctly classified, while two samples were unknown classified. One of these was the *Soleta* sample, labeled as missing in **Figure 2B**, showing the model goodness against misclassifications of new almond varieties.

MS/MS experiments were acquired for these 20 ions and searched in online databases (*METLIN* or *HMDB*). When no results were obtained (12 out of 20), they were tried to be elucidated with MetFrag in-silico fragmentation tool, annotating only 2 additional compounds and still remaining 10 ions to be elucidated manually, the most complex and lengthy step in the metabolomics workflow. In this case, we have tentatively elucidated four extra markers leaving six markers only as chemical formula.

M268T85, M450T190, M348T852, M166T99, M318T239, M503T55, M377T68 and M577T127 were found in METLIN and tentatively elucidated after comparing their experimental spectra with online available. M287T340 and M593T393 were tentatively elucidated with MetFrag tool, selecting the highest scoring molecules. For the rest of the compounds, M464T246, M548T340, M647T403 and M476T114 were finally manually elucidated. An example of manual elucidation is shown **in Figure 3**. MS/MS experiments were carried out for M464T246 marker, annotated as an ammonium adduct, based on accurate mass full scan spectra (**Figure S3**). A product ion at *m/z* 285.1343 (+1.1 mDa mass error) was observed at 10eV corresponding to the loss of NH_3 plus $C_6H_{10}O_5$ group. This pointed out the presence of at least one hexose unit in the molecule. Then, two consecutive water losses were observed at *m/z* 267.1216 (+0.3 mDa) and 249.1114 (+0.3 mDa). Furthermore, product ion at *m/z* 163.0602 was assigned to an additional hexose unit showing an elemental composition $[C_6H_{11}O_5]^+$, which is supported by two other consecutive water losses at *m/z* 145.0501 (-0.4 mDa) and 127.0387 (0.9 mDa). At this point, this compound was tentatively elucidated as hexosyl-2-phenylethyl glucopyranoside as shown in **Table 2**. In the same way, the rest of the compounds were tried to be elucidated. Six of them were not tentatively elucidated as more than one compound fit with the experimental spectra, anyway the two main product ions are reported in **Table S2**. With these 20 compounds, the 95% of the samples were correctly classified regarding their variety, as the Soleta

sample was not misclassified, showing the robustness of the model to assess the correct variety and to avoid false positive assignations.

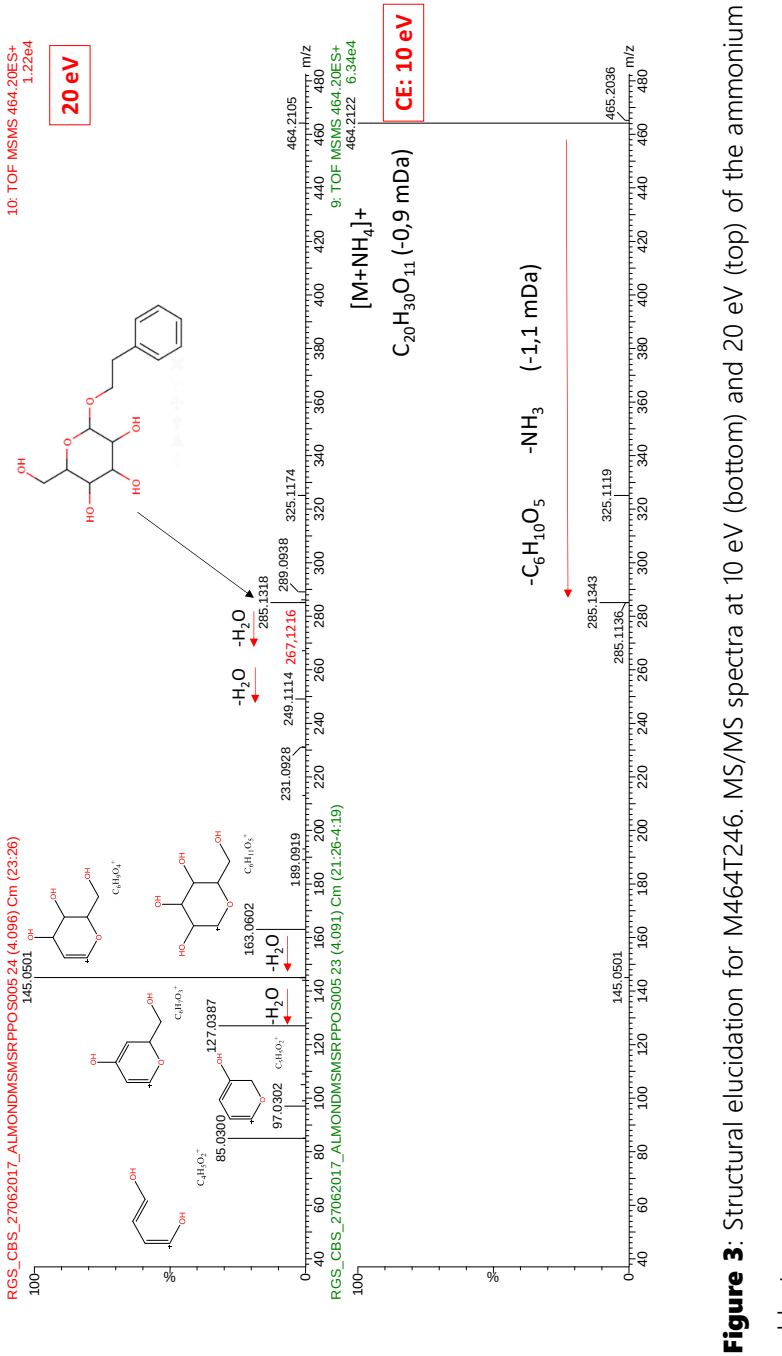


Figure 3: Structural elucidation for M464T246. MS/MS spectra at 10 eV (bottom) and 20 eV (top) of the ammonium adduct.

Conclusions

This work has shown that untargeted metabolomics is a powerful technique to develop classification models to differentiate food, based not only on their origin but also on variety. The selected extraction procedure provides a fast and easy analysis, obtaining robust results. Additionally, the power of UHPLC-HRMS technique allows to analyze a wide range of low-concentrated compounds, highlighting the most differentiating ones. The analysis of both positive and negative ionization modes gives information of a wider range of acidities, which supported by the appropriate HRMS sensitivity, highlights the best compounds to create classification models.

One of the advantages of QTOF instruments is the possibility to perform tandem mass spectrometry experiments with accurate mass information, which strongly helps in the elucidation process. The model has allowed to discriminate the origin of the almonds with only 5 compounds, ensuring the differentiation between Spanish and USA almonds, also avoiding the inclusion of bitter almonds. Furthermore, 20 marker have been selected to discriminate the almond variety, obtaining 95% of correctly classified samples.

Even though, these promising results will be validated with a larger sample set, ensuring that the models continue being robust and accurate in following seasons.

Acknowledgments

Authors acknowledge the support from Generalitat Valenciana (Group of Excellence Prometeo II/2017/023). This work has also been developed with financial support from Universitat Jaume I (UJI-B2016-10).

References

Alasalvar, C., & Bolling, B. W. (2015). Review of nut phytochemicals, fat-soluble bioactives, antioxidant components and health effects. *British Journal of Nutrition*, 113(S2), S68–S78.

Amorello, D., Orecchio, S., Pace, A., & Barreca, S. (2016). Discrimination of almonds (*Prunus dulcis*) geographical origin by minerals and fatty acids profiling. *Natural Product Research*, 30(18), 2107–2110.

Barreira, J. C. M., Casal, S., Ferreira, I. C. F. R., Peres, A. M., Pereira, J. A., & Oliveira, M. B. P. P. (2012). Supervised Chemical Pattern Recognition in Almond (*Prunus dulcis*) Portuguese PDO Cultivars: PCA- and LDA-Based Triennial Study. *Journal of Agricultural and Food Chemistry*, 60(38), 9697–9704.

Beltrán, E., Ibáñez, M., Portolés, T., Ripollés, C., Sancho, J. V., Yusà, V., … Hernández, F. (2013). Development of sensitive and rapid analytical methodology for food analysis of 18 mycotoxins included in a total diet study. *Analytica Chimica Acta*, 783, 39–48.

Beltrán Sanahuja, A., Ramos Santonja, M., Grané Teruel, N., Martín Carratalá, M. L., & Garrigós Selva, M. C. (2011). Classification of Almond Cultivars Using Oil Volatile Compound Determination by HS-SPME–GC–MS. *Journal of the American Oil Chemists' Society*, 88(3), 329–336.

Bullo, M., Lamuela-Raventos, R., & Salas-Salvado, J. (2011). Mediterranean Diet and Oxidation: Nuts and Olive Oil as Important Sources of Fat and Antioxidants. *Current Topics in Medicinal Chemistry*, 11(14), 1797–1810.

Castro-Puyana, M., & Herrero, M. (2013). Metabolomics approaches based on mass spectrometry for food safety, quality and traceability. *TrAC Trends in Analytical Chemistry*, 52, 74–87.

Cevallos-cevallos, J. M., Etxeberria, E., Danyluk, M. D., & Rodrick, G. E. (2009). Metabolomic analysis in food science : a review. *Trends in Food Science & Technology*, 20(11–12), 557–566.

Emwas, A.-H. M. (2015). The strengths and weaknesses of NMR spectroscopy and mass spectrometry with particular focus on metabolomics research. *Methods in Molecular Biology* (Clifton, N.J.), 1277, 161–93.

Gil-Solsona, R., Raro, M., Sales, C., Lacalle, L., Diaz, R., Ibañez, M., … Hernández, F. J. (2016). Metabolomic approach for Extra virgin olive oil origin discrimination making use of ultra-high performance liquid chromatography - Quadrupole time-of-flight mass spectrometry. *Food Control*, 70, 350–359.

Hyson, D. A., Schneeman, B. O., & Davis, P. A. (2002). Almonds and almond oil have similar effects on plasma lipids and LDL oxidation in healthy men and women. *The Journal of Nutrition*, 132(4), 703–7.

Jamshed, H., Gilani, A.-H., Sultan, F. A. T., Amin, F., Arslan, J., Ghani, S., … Franco, M. (2015). Almond supplementation reduces serum uric acid in coronary artery disease patients: a randomized controlled trial. *Nutrition Journal*, 15(1), 77.

Kodad, O., & Socias i Company, R. (2008). Variability of Oil Content and of Major Fatty Acid Composition in Almond (Prunus amygdalus Batsch) and Its Relationship with Kernel Quality. *Journal of Agricultural and Food Chemistry*, 56(11), 4096–4101.

Petroselli, G., Mandal, M. K., Chen, L. C., Hiraoka, K., Nonami, H., & Erra-Balsells, R. (2015). In situ analysis of soybeans and nuts by probe electrospray ionization mass spectrometry. *Journal of Mass Spectrometry*, 50(4), 676–682.

Riedl, J., Esslinger, S., & Fauhl-Hassek, C. (2015). Review of validation and reporting of non-targeted fingerprinting approaches for food authentication. *Analytica Chimica Acta*, 885, 17–32.

Romero, A. (2014). Almond quality requirements for industrial purposes - its relevance for the future acceptance of new cultivars from breeding programs. *Acta Horticultura*e, (1028), 213–220.

Ros, E., & Emilio. (2010). Health Benefits of Nut Consumption. *Nutrients*, 2(7), 652–682.

Rubert, J., Zachariasova, M., & Hajslova, J. (2015). Advances in high-resolution mass spectrometry based on metabolomics studies for food – a review. *Food Additives & Contaminants: Part A,* 32(10), 1685–1708.

Ruiz-Aracama, A., Lommen, A., Huber, M., Van De Vijver, L., & Hoogenboom, R. (2011). Application of an untargeted metabolomics approach for the identification of compounds that may be responsible for observed differential effects in chickens fed an organic and a conventional diet. *Food Additives & Contaminants*: *Part A*, 1–10.

Sales, C., Cervera, M. I., Gil, R., Portol??s, T., Pitarch, E., & Beltran, J. (2017). Quality classification of Spanish olive oils by untargeted gas chromatography coupled to hybrid quadrupole-time of flight mass spectrometry with atmospheric pressure chemical ionization and metabolomics-based statistical approach. *Food Chemistry*, 216, 365–373.

Shen, Q., Dong, W., Yang, M., Li, L., Cheung, H. Y., & Zhang, Z. (2013). Lipidomic fingerprint of almonds (*prunus dulcis* L. cv nonpareil) using TiO2 nanoparticle based matrix solid-phase dispersion and MALDI-TOF/MS and its potential in geographical origin verification. Journal of Agricultural and *Food Chemistry*, 61(32), 7739–7748.

Smith, C. A., Want, E. J., O'Maille, G., Abagyan, R., & Siuzdak, G. (2006). XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal Chem*, 78(3), 779–787.

Ukhanova, M., Wang, X., Baer, D. J., Novotny, J. A., Fredborg, M., Mai, V., … Canani, R. B. (2014). Effects of almond and pistachio consumption on gut microbiota composition in a randomised cross-over human feeding study. *British Journal of Nutrition*, 111(12), 2146–2152.

Yada, S., Lapsley, K., & Huang, G. (2011, June 1). A review of composition studies of cultivated almonds: Macronutrients and micronutrients. *Journal of Food Composition and Analysis. Academic Press.*

SUPPLEMENTARY MATERIAL

Table S1: Main product ions for the selected markers in origin model.

Feature	Molecular formula	Ionization mode	Accurate mass (error) Parent ion	Retention time (min)	Accurate mass (error, CE) neutral loss Product ion 1	Accurate mass (error, CE) neutral loss Product ion 2
M448T119	$C_{18}H_{25}NO_{12}$	POSITIVE	448.1458 (0.3 mDa)	1.99	286.0927 (0.0 mDa, 10eV) $C_{12}H_{16}NO_7$	124.0415 (1.6 mDa, 20eV) $C_6H_6NO_2$
M298T178	$C_{11}H_{15}N_5O_5S$	POSITIVE	298.0959 (-1.5 mDa)	2.98	136.0622 (-0.1 mDa, 20eV) $C_5H_6N_5$	119.0339 (-1.9 mDa, 40eV) $C_5H_3N_4$
M293T201	$C_{12}H_{22}O_8$	NEGATIVE	293.1231 (-0.5 mDa)	3.35	131.0689 (-1.9 mDa, 20eV) $C_6H_{11}O_3$	101.0229 (-1.0 mDa, 20eV) $C_4H_5O_3$
M318T239	$C_{14}H_{27}NO_5$	POSITIVE	318.2015 (-1.4 mDa)	3.94	219.1333 (-1.2 mDa, 20eV) $C_9H_{19}N_2O_4$	72.0810 (-0.3 mDa, 40eV) $C_4H_{10}N$
M933T337	$C_{40}H_{53}NO_{23}$	POSITIVE	933.3323 (-2.8 mDa)	4.98	459.1505 (+0.2 mDa, 20eV) $C_{20}H_{27}O_{12}$	297.0948 (-2.6 mDa, 30eV) $C_{14}H_{17}O_7$

CE: Collision Energy

Table S2: Main product ions for the selected markers in variety model

Feature	Molecular formula	Ionization mode	Accurate mass (error) Parent ion	Retention time (min)	Accurate mass (error, CE) Product ion 1	Accurate mass (error, CE) Product ion 2
M503T55	$C_{18}H_{32}O_{16}$	NEGATIVE	503.1612 (+0.4 mDa)	1.00	221.0643 (-1.8 mDa, 20eV) $C_8H_{13}O_7$	179.0545 (-1.1 mDa, 30eV) $C_8H_{11}O_6$
M377T68	$C_{12}H_{22}O_{11}$	NEGATIVE	377.0851 (-1.4 mDa)	1.23	179.0546 (-1.0 mDa, 20eV) $C_6H_{11}O_6$	101.0236 (-0.3 mDa, 20eV) $C_4H_5O_3$
M268T85	$C_{10}H_{13}N_5O_4$	POSITIVE	268.1056 (-1.1 mDa)	1.46	136.0626 (+0.6 mDa, 20eV) $C_5H_6N_5$	119.0361 (+0.3 mDa, 30 eV) $C_5H_3N_4$
M166T99	$C_6H_{11}NO_2$	POSITIVE	166.0868 (-0.1 mDa)	1.64	120.0815 (+0.2 mDa, 10eV) $C_8H_{10}N$	103.0551 (+0.3 mDa, 30eV) C_8H_7
M476T114	$C_{20}H_{29}NO_{12}$	POSITIVE	476. 1768 (+1.8 mDa)	1.79	314.1224 (-1.6 mDa, 20eV) $C_{14}H_{20}NO_7$	152.0711 (-0.1 mDa, 20eV) $C_8H_{10}NO_2$
M577T127	$C_{30}H_{26}O_{12}$	NEGATIVE	577.1346 (-1.2 mDa)	2.05	425.0846 (-2.7 mDa, 10eV) $C_{22}H_{17}O_9$	289.0709 (-0.3 mDa, 10eV) $C_{15}H_{13}O_6$
M450T190	$C_{19}H_{38}O_{11}$	POSITIVE	450.1975 (+2.7 mDa)	3.14	325.1097 (-3.8 mDa, 20eV) $C_{12}H_{21}O_{10}$	145.0488 (-1.3 mDa, 20eV) $C_6H_9O_4$
M313T217	$C_{14}H_{17}NO_6$	POSITIVE	313.1372 (+2.2 mDa)	3.58	180.0873 (+0.1 mDa, 10eV) $C_6H_{14}NO_5$	145.0503 (-2.5 mDa, 20eV) C_9H_7NO
M289T214	$C_{15}H_{14}O_6$	NEGATIVE	289.0712 (-0.6 mDa)	3.6	245.0797 (-1.7 mDa, 20eV) $C_{14}H_{13}O_4$	123.0444 (-0.2 mDa, 30eV) $C_7H_7O_2$
M318T239	$C_{14}H_{27}N_3O_5$	POSITIVE	318.2015 (-1.4 mDa)	3.94	219.1333 (-1.2 mDa, 20eV) $C_9H_{19}N_2O_4$	72.0810 (-0.3 mDa, 40Ev) $C_4H_{10}N$
M464T246	$C_{20}H_{32}O_{11}$	POSITIVE	464.2131 (-0.9 mDa)	4.09	285.1343(+0.5 mDa, 20eV) $C_{14}H_{21}O_6$	145.0501(0.0 mDa, 20eV) $C_6H_9O_4$
M548T340	$C_{24}H_{36}O_{13}$	POSITIVE	548.2343 (+1.2 mDa)	5.65	369.1543 (-0.6 mDa, 10eV) $C_{18}H_{25}O_8$	207.1019 (-0.2 mDa, 20eV) $C_{12}H_{15}O_3$
M563T342	$C_{28}H_{39}N_3O_{12}$	POSITIVE	563.1877 (+0.5 mDa)	5.7	430.1350 (+0.1 mDa, 10eV) $C_{18}H_{24}NO_{11}$	268.0817 (-0.4 mDa, 20eV) $C_{12}H_{14}NO_6$
M540T362	$C_{26}H_{36}O_{11}$	POSITIVE	540.2444 (+0.2 mDa)	6.05	345.1692 (-1.0 mDa, 10eV) $C_{20}H_{25}O_5$	221.1175 (-0.3 mDa, 20eV) $C_{13}H_{17}O_3$
M521T362	$C_{26}H_{34}O_{11}$	NEGATIVE	521.1983 (+3.0 mDa)	6.05	341.1385 (-0.4 mDa, 10eV) $C_{20}H_{21}O_5$	179.0552 (-0.4 mDa, 10eV) $C_8H_{11}O_6$
M287T340	$C_{15}H_{12}O_6$	NEGATIVE	287.0556 (-1.4 mDa)	6.07	259.0595 (-1.1 mDa, 10eV) $C_{14}H_{11}O_5$	125.0232 (-0.7 mDa, 20eV) $C_6H_5O_3$
M593T393	$C_{27}H_{30}O_{15}$	NEGATIVE	593.1506 (-0.1 mDa)	6.56	285.0381 (-1.8 mDa, 30eV) $C_{15}H_9O_6$	255.0290 (-0.3 mDa, 30eV) $C_{14}H_7O_5$
M477T395	$C_{22}H_{22}O_{12}$	NEGATIVE	477.1033 (-0.7 mDa)	6.61	314.0417 (-1.0 mDa, 20eV) $C_{16}H_{10}O_7$	243.0284 (-0.9 mDa, 20eV) $C_{13}H_7O_5$
M647T403	$C_{28}H_{32}O_{16}$	POSITIVE	642.2034 (0.0 mDa)	6.72	479.1187 (-0.3 mDa, 10eV) $C_{22}H_{22}O_{12}$	317.0658 (-0.3 mDa, 20eV) $C_{16}H_{13}O_7$
M348T852	$C_{20}H_{42}O_3$	POSITIVE	348.3477 (-1.0 mDa)	14.06	163.1337 (+0.3 mDa, 10eV) $C_8H_{19}O_3$	89.0634 (+3.1 mDa, 20eV) $C_4H_9O_2$

CE: Collision Energy

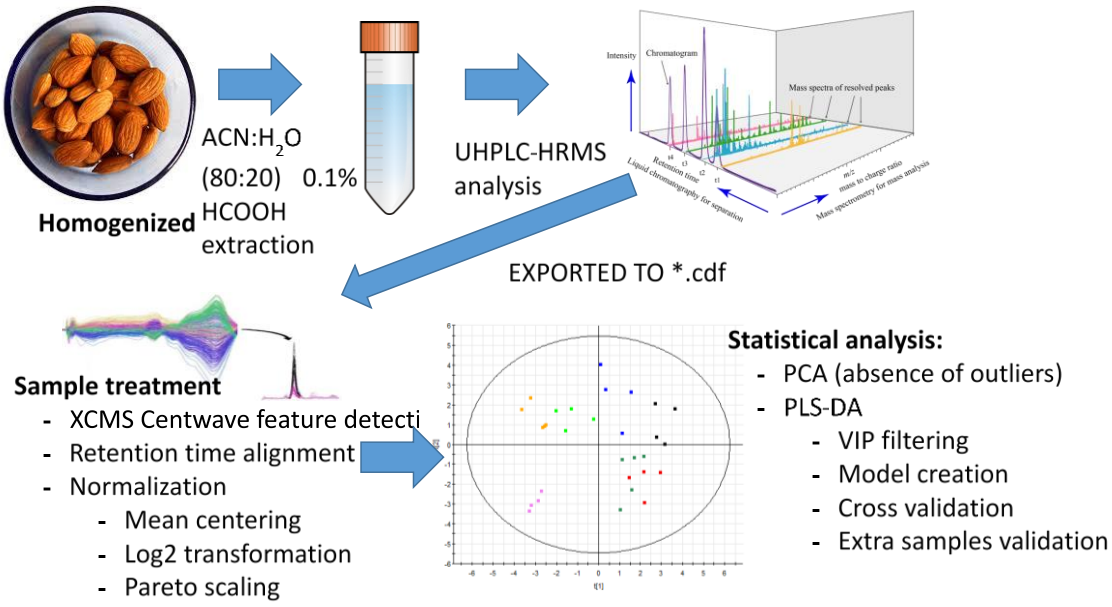


Figure S1: Sample and data treatment workflow for untargeted metabolomics studies

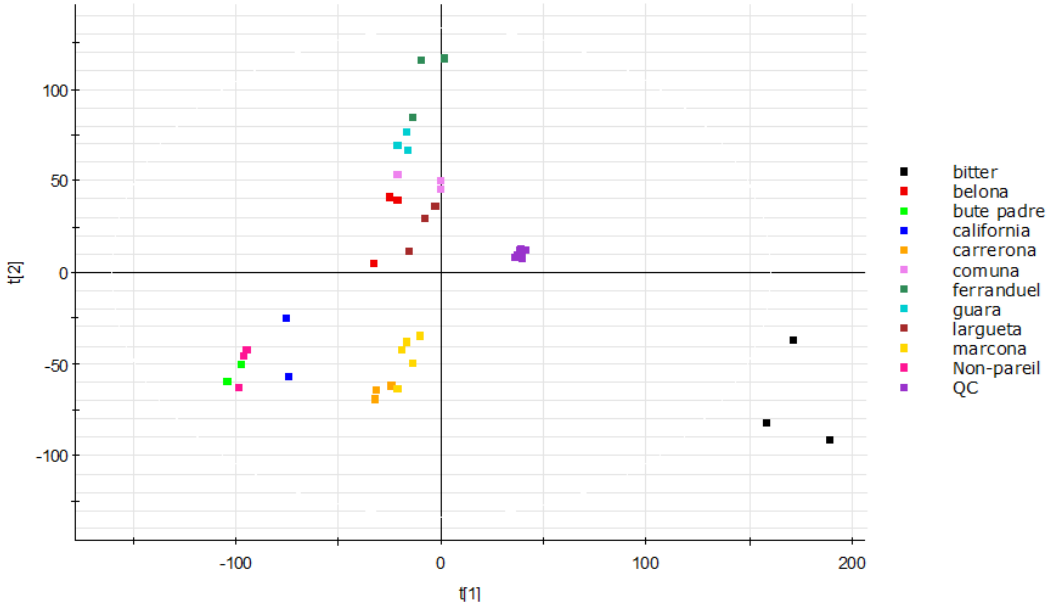


Figure S2: PCA score plot of model samples.

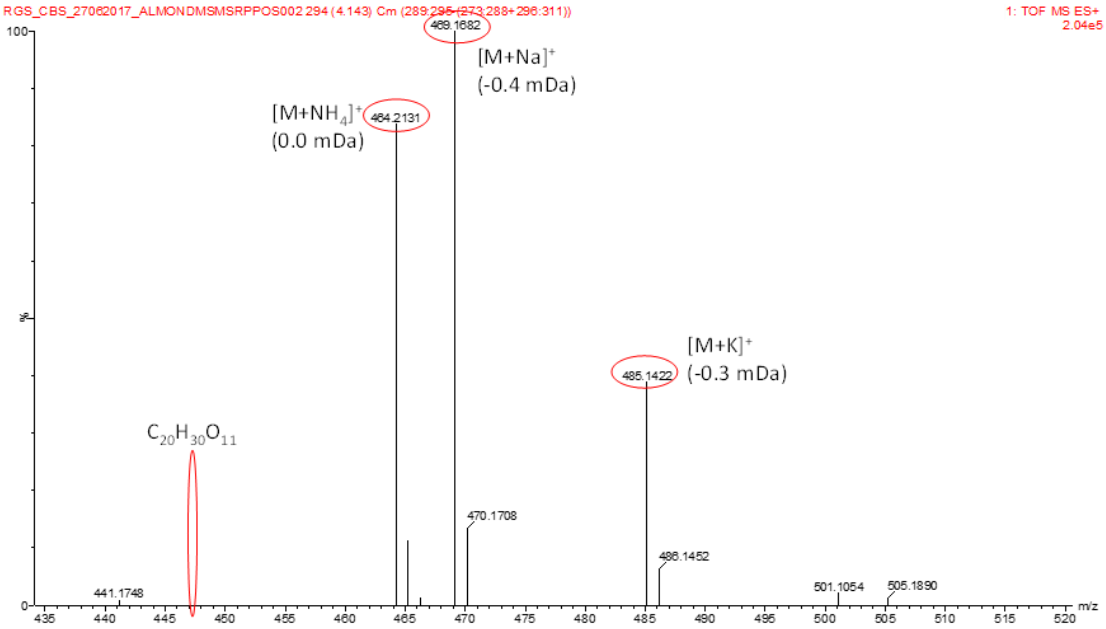


Figure S3: Low Energy spectrum for M464T246 chromatographic peak observing the ammonium, sodium and potassium adducts.

III.3: Artículo científico 5

Food Chemistry 250 (2018) 89–97

Contents lists available at ScienceDirect

Food Chemistry

journal homepage: www.elsevier.com/locate/foodchem

Assessment of protected designation of origin for Colombian coffees based on HRMS-based metabolomics

Duvan E. Hoyos Ossa[a,1], Rubén Gil-Solsona[b,1], Gustavo A. Peñuela[a], Juan Vicente Sancho[b,*], Felix J. Hernández[b]

[a] GIDCON Research Group, Faculty of Engineering, University Research Headquarters (SIU), University of Antioquia, Street 70 # 52 – 21, Medellín, Colombia
[b] Research Institute for Pesticides and Water (IUPA), Avda. Sos Baynat, s/n. University Jaume I, 12071 Castellón, Spain

ARTICLE INFO

Keywords:
Colombian coffee
Protected designation of origin
Liquid chromatography
High resolution mass spectrometry
Quadrupole time-of-flight
Metabolomics

ABSTRACT

An untargeted metabolomics approach based on HRMS has been applied to Colombian green coffee to develop a discrimination model to highlight the most differential compounds. For this purpose, 41 green coffee samples of different genotypes collected from 5 regions were analysed. Samples were extracted with aqueous and organic solvents to cover a wide range of compounds. Sample extracts were randomly injected and data were pre-processed with XCMS software. PCA was used to verify quality control samples behaviour, and PLS-DA and so-SIMCA were employed to create models for discrimination using VIP variable selection method. Thirteen different compounds correctly separate green coffee samples according to their origin, several related to the quality and health benefits of coffee. Model validation was achieved using both cross-validation and an additional set with coffee samples from different harvest year. The results reveal that UHPLC-(Q)ToF MS-based metabolomics is a suitable tool to develop food origin discrimination strategies.

1. Introduction

Colombia is one of the most important coffee producers, exporting about 7% premium coffee to the rest of the world between 2000 and 2011, which represents 9% Colombian exports (García, Posada-Suárez, & Liderach, 2014). From 2007, 'Café de Colombia' has been registered as Protection of Geographical Indications granted by the European Commission, like a recognition for its high quality and characteristics like mild, clean cup, medium/high acidity and body, and a full of pronounced aroma. The main coffee varieties of Arabica species cultivated in Colombia are Borbón, Castillo, Caturra, Colón, Tabi, Típica, Maragogipe and San Bernardo.

Protected Designation of Origin (PDO) and Protection of Geographical Indications (PGI) have a specific link to the region where the product comes from, according to its quality and specific production conditions. The quality and reputation of a PDO/PGI product such as *'Café de Colombia'* are explained by natural factors like weather, soil, altitude, as well as human knowledge and tradition. In Colombia, *'La Superintendencia de Industria y Comercio – SIC'* has granted PDO recognition for the products *'Café de Colombia (2005)'*, *'Café de Cauca (2011)'*, *'Café de Nariño (2011)'*, *'Café del Huila (2013)'*, *'Café de Santander (2014)'* and recently *'Café de la Sierra Nevada (2017)'* and *'Café del Tolima (2017)'*.

It is important to ensure the origin of this coffee for producer's benefits and to protect consumers from adulterations and/or misrepresentations. To this aim, research has been carried out using Nuclear Magnetic Resonance spectroscopy (NMR) (Arana et al., 2015; Wei et al., 2012), Near-Infrared spectroscopy (Oberthür et al., 2011) and Gas Chromatography with several detection systems (Arana, Medina, Esseiva, Pazos, & Wist, 2016). Different biomarkers were found for Colombian coffee classification faced to other countries. However, it is also important to ensure geographical zone origin inside the country due to the differences in coffee composition (Villarreal et al., 2009) and/or product taste. In order to certify this origin, PDO was created to provide consumers confidence about correct origin assurance. Recent studies have been carried out, obtaining biomarkers based on chemical signals directly related to the origin (Oberthür et al., 2011), as a response to the interest of ensuring PDO for these premium products. In the case of Colombian coffee, The National Coffee Research Center-CENICAFÉ plays an important role to ensure the origin and to enhance coffee quality and traceability of Colombian coffee. Despite the coffee variety cultivated in Colombia is Arabica, gene-modifications have been developed to fight against diseases (Castillo, Cristancho, Isaza, Pinzón, & Rodríguez, 2014). These modifications may affect Protected

* Corresponding author.
E-mail address: sancho@uji.es (J.V. Sancho).
[1] Both authors contributed equally to this work and are co-first authors.

https://doi.org/10.1016/j.foodchem.2018.01.038
Received 13 September 2017; Received in revised form 18 December 2017; Accepted 3 January 2018
Available online 04 January 2018
0308-8146/

Assessment of protected designation of origin for Colombian coffees based on HRMS-based metabolomics

Duvan E. Hoyos Ossa [*a], Rubén Gil-Solsona [*b], Gustavo A. Peñuela [a], Juan Vicente Sancho [**b], Felix J. Hernández [b]

[a] GDCON Research Group, Faculty of Engineering, University Research Headquarters (SIU), University of Antioquia, Street 70 # 52 - 21, Medellin, Colombia.

[b] Research Institute for Pesticides and Water (IUPA). Avda. Sos Baynat, s/n. University Jaume I, 12071 Castellón, Spain.

* Both authors contributed equally to this work and are co-first authors

** Corresponding author

Abstract

An untargeted metabolomics approach based on HRMS has been applied to Colombian green coffee to develop a discrimination model to highlight the most differential compounds. For this purpose, 41 green coffee samples of different genotypes collected from 5 regions were analysed. Samples were extracted with aqueous and organic solvents to cover a wide range of compounds. Sample extracts were randomly injected and data were pre-processed with XCMS software. PCA was used to verify quality control samples behaviour, and PLS-DA and DD-SIMCA were employed to create models for discrimination using VIP variable selection method. Thirteen different compounds correctly separate green coffee samples according to their origin, several related to the quality and health benefits of coffee. Model validation was achieved using both cross-validation and an additional set with coffee samples from different harvest year. The results reveal that UHPLC-(Q)ToF MS-based metabolomics is a suitable tool to develop food origin discrimination strategies.

Introduction

Colombia is one of the most important coffee producers, exporting about 7 % premium coffee to the rest of the world between 2000 and 2011, which represents 9 % Colombian exports (*García L, Posada-Suárez, & Läderach, 2014*). From 2007, 'Café de Colombia' has been registered as Protection of Geographical Indications granted by the European Commission, like a recognition for its high quality and characteristics like mild, clean cup, medium/high acidity and body, and a full of pronounced aroma. The main coffee varieties of Arabica species cultivated in Colombia are *Borbón, Castillo, Caturra, Colón, Tabi, Típica, Maragogipe* and *San Bernardo*.

Protected Designation of Origin (PDO) and Protection of Geographical Indications (PGI) have a specific link to the region where the product comes from, according to its quality and specific production conditions. The quality and reputation of a PDO/PGI product such as 'Café de Colombia' are explained by natural factors like weather, soil, altitude, as well as human knowledge and tradition. In Colombia, 'La Superintendencia de Industria y Comercio – SIC' has granted PDO recognition for the products 'Café de Colombia (2005)', 'Café de Cauca (2011)', 'Café de Nariño (2011)', 'Café del Huila (2013)', 'Café de Santander (2014)' and recently 'Café de la Sierra Nevada (2017)' and 'Café del Tolima (2017)'.

It is important to ensure the origin of this coffee for producer's benefits and to protect consumers from adulterations and/or misrepresentations. To this aim, research has been carried out using Nuclear Magnetic Resonance spectroscopy (NMR) (*V A Arana et al., 2015; Wei et al., 2012*), Near-Infrared spectroscopy (*Oberthür et al., 2011*) and Gas Chromatography with several detection systems (*Victoria Andrea Arana, Medina, Esseiva, Pazos, & Wist, 2016*). Different biomarkers were found for Colombian coffee classification faced to other countries. However, it is also important to ensure geographical zone origin inside the country due to the differences in coffee composition (Villarreal et al., 2009) and/or product taste. In order to certify this origin, PDO was created to provide consumers confidence about correct origin assurance. Recent studies have been carried out, obtaining biomarkers based on chemical signals directly related to the origin (*Oberthür et al., 2011*), as a response to the interest of ensuring PDO for these premium products. In the case of Colombian coffee, The National Coffee Research Center-CENICAFÉ plays an important role to ensure the origin

and to enhance coffee quality and traceability of Colombian coffee. Despite the coffee variety cultivated in Colombia is Arabica, gene-modifications have been developed to fight against diseases (*Castillo, Cristancho, Isaza, Pinzón, & Rodríguez, 2014*). These modifications may affect Protected Denomination of Origin (PDO) markers, which increases the interest of studying this fact.

In last few years, research efforts have been directed towards the identification of PDO chemomarkers for premium food products, such as Extra Virgin Olive Oil (*Gil-Solsona et al., 2016; Sales et al., 2017*), Saffron (*Rubert, Lacina, Zachariasova, & Hajslova, 2016*), Spanish 'Jamón de Teruel' (*Fandos-Herrera, 2016*) , Vinegars (*Chinnici, Durán-Guerrero, & Riponi, 2015*) , Wines (*Díaz et al., 2016*) or Brazilian coffee (*Nagai, Santini Pigatto, & Lourenzani, 2016*). In the field of coffee authentication, studies have been reported for variety discrimination (*Montero-Vargas et al., 2013*) as well as specialty processing like Asian palm civet coffee (*Jumhawan et al., 2013*) due to its high market prize and taste. Regarding coffee origin discrimination different mass spectrometric techniques have been used as Isotope ratio mass spectrometry [IRMS] (*Rodrigues et al., 2011; Serra et al., 2005; Valentin & Watling, 2013; Weckerle, Richling, Heinrich, & Schreier, 2002*), LC-MS (*Alonso-Salces, Serra, Reniero, & Héberger, 2009*) or GC-MS (*Risticevic, Carasek, & Pawliszyn, 2008*). In the specific case of Colombian coffee, for its differentiation from other coffees from the rest of the world additional spectroscopic techniques such as 1H-NMR has also been used (*Victoria Andrea Arana et al., 2016*). However, due to the valuable information provided by high resolution MS in the identification of individual markers, a special effort should be made to implement a HRMS method for Colombian coffees PDO discrimination.

Metabolomics has been boosted as an important tool to highlight biomarkers (*Castro-Puyana & Herrero, 2013; Dettmer, Aronov, & Hammock, 2007*). The combination of Ultra-High Performance Liquid Chromatography (UHPLC) coupled to a hybrid Quadrupole Time-Of-Flight Mass Spectrometer (Q-Tof MS) merges an excellent separation technique with accurate mass measurements, enhancing the identification power of this technique for compounds of different polarities and over a wide concentration range (*Gallart-Ayala, Chéreau, Dervilly-Pinel, & Le Bizec, 2015*). In this work, an analytical strategy has been developed to classify different genotypes derived from Caturra x Timor hybrid (CIFC 1343) components of Castillo variety, cultured at different sites of the Colombian coffee region, in order to highlight the most robust biomarkers for Colombian PDO assessment, evaluating how the genotypes and the farming sites affects to PDO biomarkers.

Materials & methods

Chemicals

HPLC-grade water was obtained by purifying demineralized water in a Mili-Q plus system from Millipore (Bedford, MA, USA). HPLC-supergradient acetonitrile (ACN), HPLC-grade methanol (MeOH), HPLC-supergradient acetone, HPLC-supergradient isopropanol (i-PrOH) and reagent-grade ammonium acetate (NH4Ac) were obtained from Scharlab (Barcelona, Spain). Leucine-enkephalin and formic acid (HCOOH, 98 - 100 %) were purchased from Sigma-Aldrich (Augsburg, Germany).

Instrumentation and equipment.

A Waters Acquity UPLC system (Waters, Milford, MA, USA) was interfaced to a hybrid Quadrupole-ToF High Resolution Mass Spectrometer (HRMS) (Xevo G2 QTof, Waters, Manchester, UK) using a Z-spray-ESI interface operating in both positive and negative ionization modes with resolution of the ToF mass spectrometer about 20000 at full width half maximum (FWHM). A MiVac Duo concentrator (Genevac, United Kingdom) was employed to led sample extracts to dryness under vacuum.

Sample cultivars and harvest

A total amount of 41 green coffee samples belonging to 10 genotypes derived from Caturra x Timor hybrid (CICF 1343) components of Castillo variety were collected in 2015 from five (5) different CENICAFE experimental stations located in the Colombian coffee territory in order to create the classification model. *Naranjal, Rosario, Gigante, Sirena* and *San Antonio* are the names for the experimental stations located in the departments of Caldas, Antioquia, Huila, Valle del Cauca and Santander respectively, all them between 1300 and 1600 meters above sea level according to the register for PGI 'Café de Colombia'. These experimental stations meet the environmental conditions representative of most of the country's coffee farms. In a second season sampling (2016), 32 green coffee samples belonging to the same experimental stations and genotypes were collected and used for model validation.

Sample treatment.

Harvested samples were stored at - 80 °C, passed through liquid nitrogen and triturated in an ultra-centrifugal mill at 14000 RPM. Two-extraction processes were applied to triturated green coffee samples in order to extract the largest number of polar and semi-polar compounds.

In the first extraction, 3.0 g of sample were extracted with 10.0 mL of acetonitrile/water/formic acid (79:20:1, v/v/v) mixture, mechanically shaken 90 min, sonicated 15 min and centrifuged at 6000 RPM during 10 min. The supernatant was divided in 2 aliquots of 2 mL each. Firstly, 2 mL of the supernatant were diluted with 6 mL of Mili-Q water and the solution was used for Reversed Phase (RP) liquid chromatography analysis in positive and negative mode. Another 2 mL aliquot was diluted with 6 mL of acetonitrile and the solution was used for Hydrophilic Interaction LIquid Chromatography (HILIC) analysis in positive and negative ionization mode. For Quality Control (QC) samples preparation, 100 µL of each extract were pooled. This mixture was injected at the beginning of the sample batch in order to equilibrate the column and every 10 samples to control the possible signal drift along the batch.

In the second extraction, 3.0 g triturated sample were extracted with 10.0 mL acetone, mechanically shaken 90 min, sonicated 15 min and centrifuged at 6000 RPM during 10 min. 2 mL of supernatant were taken and evaporated to dryness under vacuum using a MiVac Duo concentrator. The residue was reconstituted with 8.0 mL ACN and it was analysed by RPLC analysis in both positive and negative ionization modes. QC samples were also prepared as described above.

All the samples were randomly injected in all batches in order to avoid instrumental drift effect over the results.

UHPLC-QTof MS analysis of green coffee extracts

Three different UHPLC separations were carried out in order to analyse the polar and semi-polar metabolites extracted from the green coffee samples. RPLC and HILIC were employed to separate the polar fraction, while the semi-polar fraction was evaluated using RPLC with a different column, as shown in **Table S1.** A 2.7 µm fused-core column (CORTECS®) was employed for analysis of the polar fraction due to the lower backpressure required in comparison with 1.7 µm column

(ACQUITY®). In the semi-polar fraction, the high viscosity of required organic solvent (butanol) involves a high backpressure even using a fused-core column. In order to keep it within limits, a temperature increase (up to 60 ºC or even higher) was necessary, making compulsory the use of more stable column like ACQUITY based on hybrid particles.

The capillary voltage of Xevo G2 Q ToF was set at 0.7 and 1.5 kV for ESI in positive and negative ionization mode respectively and a cone voltage of 25 V were used. The source temperature was set at 120 °C and desolvation gas (N2) temperature was set at 500 °C (250 °C for semi-polar fraction) with a flow of 800 L/h and a cone gas flow of 80 L/h. MS data were acquired over an *m/z* range of 50 – 1200 Da. Argon was used as collision gas (Purity 99.995 %, Praxair, Valencia, Spain). For MSE experiments, two acquisition functions with different collision energies were configured: the low energy (LE) function, selecting as collision energy 4 eV, and the high energy (HE) function, with a collision energy ramp from 15 to 40 eV.

External calibrations were conducted from *m/z* 50 to 1200 Da with a 1:1 mixture of 0.05 M NaOH: 5 % (v/v) HCOOH diluted (1:25) with H_2O:ACN (20:80 v/v), at a flow rate of 10 µL/min. For internal lock mass calibration, a Leucine-Enkephalin solution (2 µg/mL) in ACN:H_2O (50:50 v/v) at 0.1 % HCOOH was pumped at 20 µL/min through the lock-spray needle and measured every 30 seconds, with a scan time of 0.4 seconds. Leucine-enkephalin, in positive ($[M+H]^+$, *m/z* 556.2771) and negative mode ($[M-H]^-$, *m/z* 554.2615) was used for recalibrating the mass axis during the injection and to ensure a robust accurate mass along the time. Equipment control and data acquisition were performed with Masslynx v. 4.1 software (Waters, USA).

Data processing, variable selection method and multivariate analysis

UHPLC-(Q)ToF MS data were converted from proprietary (.raw, Waters) to machine-independent data format (.cdf, NetCDF) using Databridge application from MassLynx and pre-processed using XCMS free R package. Files for positive and negative ionization modes were pre-processed separately.

For peak picking, centWave feature selection algorithm was employed, considering peak width ranging from 4 to 20 seconds, with at least 3 scans above 6000 counts, a signal to noise ratio

of 10 and 15 ppm of mass tolerance. Peak grouping (bandwidth from 15 down to 0.4 seconds) was performed to match detected features across samples before peak alignment step using the retcor() function. In order to create a list with all these features as well as to obtain peak areas, the function fillPeaks() was used. Peaks were labeled as MXXXTYYY, with XXX referring to its nominal mass and YYY to the corrected retention time in seconds. Metabolite features are defined as ions with unique *m/z* and retention time values. A LOESS (locally weighted scatter plot smoothing) normalization method was applied to the samples, in order to correct instrumental drift and log2 transformation was also used to reduce heteroscedasticity (*K. A. Veselkov et al., 2011*).

Grouping of XCMS features list was performed using CAMERA package. It implements a set of algorithms, like fast retention time-based grouping or graph-based algorithm, to integrate the peak shape, analyse the isotopic information and the intensity correlation across the samples. CAMERA was employed to avoid the use of different ions corresponding to the same compound in the model. Autoscaling (unit variance scaling) was applied to imported data in order to equilibrate the weights of all metabolites (*Di Guida et al., 2016; van den Berg, Hoefsloot, Westerhuis, Smilde, & van der Werf, 2006*). Missing value imputation was not used in the present work. Instead, zeros were replaced by N.A argument.

Multivariate data analysis was carried out using EZInfo and SIMCA-P+ software (*Umetrics, Sweden*), PLS toolbox (*Eigenvector Research Inc, US*) and DD-SIMCA GUI (*Zontov, Rodionova, Kucheryavskiy, & Pomerantsev, 2017*) within MatLab (*MathWorks, US*). Principal Component Analysis (PCA) was applied only to check the behavior of QC samples, ideally clustered in the center of the Score Plot. After that, PLS-DA was used to highlight variables (features) with potential to distinguish samples according their geographical origin (by VIP variable selection method usig *EZInfo*) and the final PLS-DA model was created using PLS Toolbox. Moreover, DD-SIMCA was also employed to create authentication models (*Rodionova, Titova, & Pomerantsev, 2016; Zontov et al., 2017*).

The order for reporting metabolomics data analysis and the vocabulary used were taken into account according to the minimum reporting standards proposed by Goodacre et al. (*Goodacre et al., 2007*).

Elucidation workflow

Accurate masses for significant markers in PLS-DA model (MXXXTYYY) were retrieved from the final XCMS table. Using the accurate mass, elemental composition calculator (MassLynx, Waters) and checking the fit of the experimental isotope distribution (i-FIT) to the theoretical one, a list for candidate molecular formula/s was obtained. MS/MS experiments at different collision energies (10, 20, 30 and 40 eV) for each marker were performed and product ion spectra employed for elucidation process.

Specific databases as *METLIN*, *Massbank*, *FooDB*, in-silico elucidation tools as *MetFrag* or candidates search in generic chemical databases as *Chemspider* were used for searching potential candidates by mean of accurate mass and/or MS/MS spectra comparison.

Validation step with second season samples

After creating the DD-SIMCA and PLS-DA models with the selected variables, the model was validated not only with a cross validation but also with a second-season samples, as recommended in the literature (*Riedl, Esslinger, & Fauhl-Hassek, 2015*). Briefly, selected compounds were manually integrated with TargetLynx (*MassLynx, Waters*) for all the samples employed in the model generation. The data was transformed (log2 transformation) and normalized with a procedure similar to probabilistic quotient normalization (PQN) described in literature (*Di Guida et al., 2016*). Instead of generating the reference vector with QC samples, it was generated with the complete set of samples, being zero (0) the minimum area across the samples and one (1) the maximum for each variable.

With these data, the PLS-DA and DD-SIMCA models were generated and the first validation was tested (cross-validation in PLS-DA and acceptance plot in DD-SIMCA). For a better validation, a "system challenge" (*Riedl et al., 2015*) was performed. Second season samples obtained from the same cultivars in a different harvesting year were injected under the same conditions, manually integrated with TargetLynx (*MassLynx, Waters*), log2 transformed and normalized in the same way as the model samples.

Then, these data were imported into the PLS-DA model and every DD-SIMCA model to observe the classification.

Results and discussion

Three different injections were made in order to cover a wide range of polarities for LC analysis. Thus, the process was initially evaluated for three green coffee injections, separating semi-polar (acetone extraction), polar (H_2O:ACN extraction injected in RPLC) and highly polar metabolites (H_2O:ACN extraction injected in HILIC).

Data pre-processing.

RPLC raw data obtained for 41 green coffee and QC samples were pre-processed using XCMS software. A total amount of 3135 (RP-), 1812 (RP+), 597 (semi-polar RP+), 412 (semi-polar RP-), 498 (HI-) and 646 (HI+) mass spectral features were extracted from selected sample batches. Retention time correction and peak integration were also carried out using XCMS software according to parameters explained before. Grouping of XCMS feature lists was performed using CAMERA [Collection of Algorithms for MEtabolite pRofile Annotation] package. Using the XCMS and CAMERA package, we could automatically group all features derived from the same compound and annotate the type of ion species (protonated molecule, sodium adduct, potassium adduct,…) in order to simplify the elucidation workflow as well as to ensure the employment of only one ion per compound.

Data pre-treatment.

Once the raw data are pre-processed, the extraction of relevant information from large data sets is a challenge in metabolomics research. Big differences in concentrations for different metabolites are not proportional to their relevance and this cannot be managed for data statistical analysis software without adequate data pre-treatment (van den Berg et al., 2006). In the case of UHPLC-HRMS metabolomics data, profiles have a heteroscedastic noise structure characterized by increasing variance as a function of increased signal intensity that can affect adversely standard statistical and pattern recognition tools (*K. A. Veselkov et al., 2011*). Due to the requirement of homoscedasticity in regression models like PLS-DA, data pre-treatment is mandatory in order to make them ready for data processing. Pre-processed data for training set was normalised using a LOESS normalization method in order to remove sources of systematic variation between samples

profiles due to factors like instrumental noise or drift in instrument detector sensitivity. Then, it was transformed using the log2 function with the aim to stabilize the technical variance: reduce heteroscedasticity (*K. A. Veselkov et al., 2011*). **Figure S1** shows the standard deviation as a function of rank (mean) signal intensity for the samples acquired in negative ionization mode in samples analysed by RPLC. Data normalization (b) was not enough to reduce heteroscedasticity although it made the raw profiles comparable in size. Finally, log2 transformation (c) removed heteroscedasticity.

Sample batch selection and processing.

To obtain repeatable and interpretable LC-MS metabolomics data, QC for each samples batch were injected (around 10) prior to samples injection, in order to stabilize the column and obtain gradual changes in instrument sensitivity over time (*K. a Veselkov et al., 2011*). QC samples, obtained by pooling equal aliquots of each extracted green coffee sample, are representative of all metabolites present in the experimental set. It is important to note that samples injection was randomized in order to affect the experimental groups by the same extent.

Principal Component Analysis (PCA), an unsupervised pattern recognition method, was applied to each training set in order to check the behaviour of the QC samples as a measure of technical variability. Once QCs grouping was verified, the checked normalized batches were transferred to PLS-DA without QCs data.

PLS-DA, a linear classification model that combines the properties of partial least squares regression with the discrimination power of a classification technique, was performed to obtain discriminant models for groups analyzed (*Ballabio & Consonni, 2013*).

A first analysis was carried out in order to separate samples by genotypes using PLS-DA within *EZInfo*. All the samples were closely similar between genotypes and subsequently the model obtained with all the features was unsuccessful (see **Figure S2**), obtaining a goodness-of-fit R^2Y=0.201 and goodness-of-prediction Q^2Y=0.068, which was lower when features were gradually reduced. Thus, we concluded that samples did not differ regarding their genotype, and therefore the genotype will not affect the origin classification of the model.

The three chromatographic separation modes were studied to select the best one to perform faster analysis. For acetone extraction injected in a RPLC C18 column (semi-polar fraction), only a goodness-of-fit R^2Y of 0.666 and goodness-of-prediction Q^2Y of 0.551, with an incorrect classification of samples in cross validation was obtained (*Szymańska, Saccenti, Smilde, & Westerhuis, 2012*). For this reason, acetone extraction was discarded to obtain the best markers.

In the case of RPLC and HILIC applied for the polar fraction analysis, a better goodness of fit was achieved. A goodness-of-fit R^2Y of 0.868 and goodness-of-prediction Q^2Y of 0.746 was obtained for RPLC analysis, and goodness-of-fit R^2Y of 0.842 and goodness-of-prediction Q^2Y of 0.751 for HILIC. Both injections provided good results in goodness of fit and prediction, making them feasible to be used to highlight the most interesting ions. Nonetheless, RPLC columns need less stabilization time, less additives for injection and are more stable against variations (pH, buffer concentrations,..). Moreover, they are more commonly employed than HILIC columns in routine laboratories. Thus, only RPLC data were finally used to build the model.

In order to select only one ionization mode, two different PLS-DA analysis were also performed in both ESI+ and ESI-. The variable importance in the projection (VIP score), that summarize the overall contribution of each X-variable to the PLS model, was used as variable ranking strategy in order to improve the training model selecting as few variables as possible and taking into account the future necessity to develop a quantitative method (*Yi et al., 2016*). Using EZ-info software, the complete VIP score list was consulted and the 50 features with higher VIP score were selected to construct the initial PLS model. Then we rebuild and validate the PLS-DA model with the 50th first features and check if good results were obtained in the validation step (more than 90% of correct classified samples). Following, the "greater-than-one-rule" was applied for the new VIP scores, and we continue until the model failed. We finally maintained the last model with >90% explained samples.

For RP- model, just 13 features were enough to obtain a good PLS-DA model. However, 30 features were even not enough to obtain good results in RP+, so the model was constructed only with the 13 best features obtained from RP- chromatography in order to simplify the model as

maximum as possible. (see **Figure S3**, where the VIP ranking for the final 13 features used for PLS-DA model construction is presented).

Next, the PLS-DA classification model was created using PLS Toolbox with these 13 ions. As can be observed in **Figure 1a**, a correct classification of all green coffee samples was achieved. However, PLS-DA models have shown problems when classifying samples belonging to classes not included in the model (the so-called alien samples), as recently reported (*Rodionova et al (2016)*. Then, additional PLS-DA models were created without the inclusion of one single class to evaluate the prediction success of the PLS-DA model against "alien" samples.

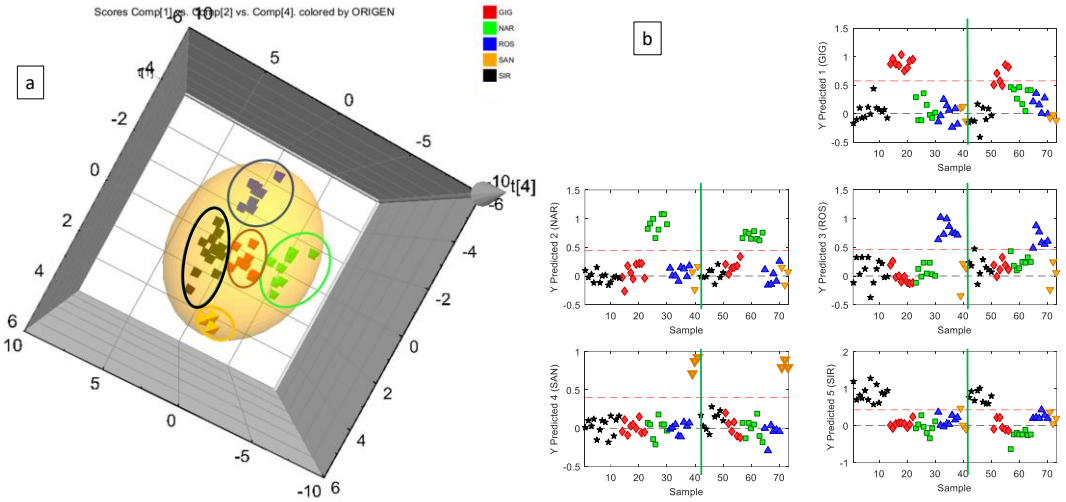


Figure 1. a) 3D score plot for components 1,2 and 4 from PLS-DA model using 13 VIP for RP-. b) Model samples classification and system challenge samples classification. Samples in the left of the green line are model samples and in the right are test samples.

In the case of removing *San Antonio* samples from the model, as shown in **Figure 2**, despite using a successful model, all *San Antonio* (alien) samples were misclassified and labelled as *Sirena* samples. The same behaviour was also observed for the rest of the classes when treated as "alien" (**Figure S4**). In order to improve the alien samples authentication, other approaches have been recently proposed, such as DD-SIMCA (*Zontov, 2017*).

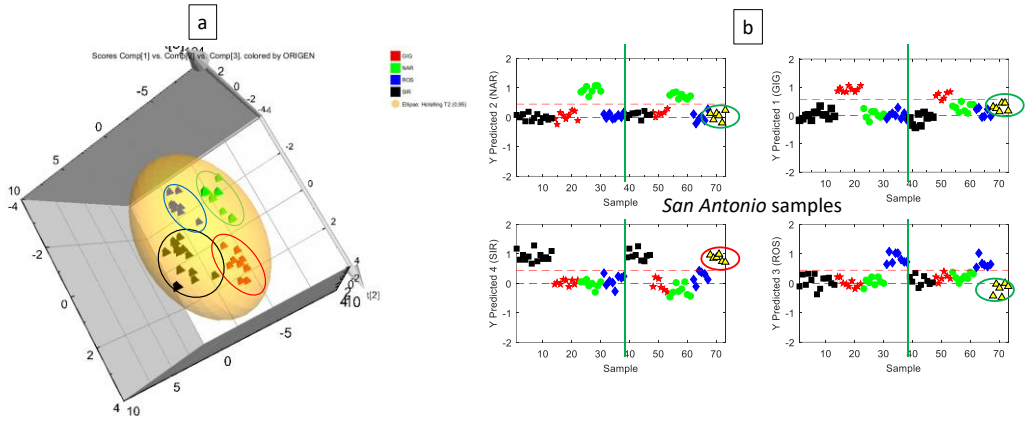


Figure 2. a) PLS-DA score plot without *San antonio* samples. b) Model samples classification (left part of the green line) and system challenge samples classification (right part of the green line). Samples above the red line are classified in the selected group. *San Antonio* samples (rounded in red) have been classified as *Sirena*.

With the same data coming from the VIP ions, five DD-SIMCA models were built per coffee class (**Figure 3** top). In this case, no misclassification was observed when building the models. The potential of DD-SIMCA will be shown in the following section involving validation step.

Model validation-prediction ability

Model validation is usually misinterpreted as simply model optimization by means of internal validation (*Goodacre et al., 2007*). For PLS-DA -a supervised method-, several statistical approaches are currently in use to validate outcomes from the model analysis, as for example cross validation procedures and permutation testing (*Szymańska et al., 2012*), usually employed to avoid the overfitting of data (*Westerhuis et al., 2008*). In our work, we used the predictive ability of the model (Q^2 and R^2) calculated by cross-validation (for which data are divided into 7 parts and each 1/7th in turn is removed to internally test it) but also the evaluation of the model with an additional validation samples set, harvested in a different year (system challenge samples).

In order to validate outcomes from the model analysis, the 13 markers selected to create the discrimination model were manually integrated in both test model and system challenge samples. A log2 transformation was applied and PQN normalization was finally employed to compare samples

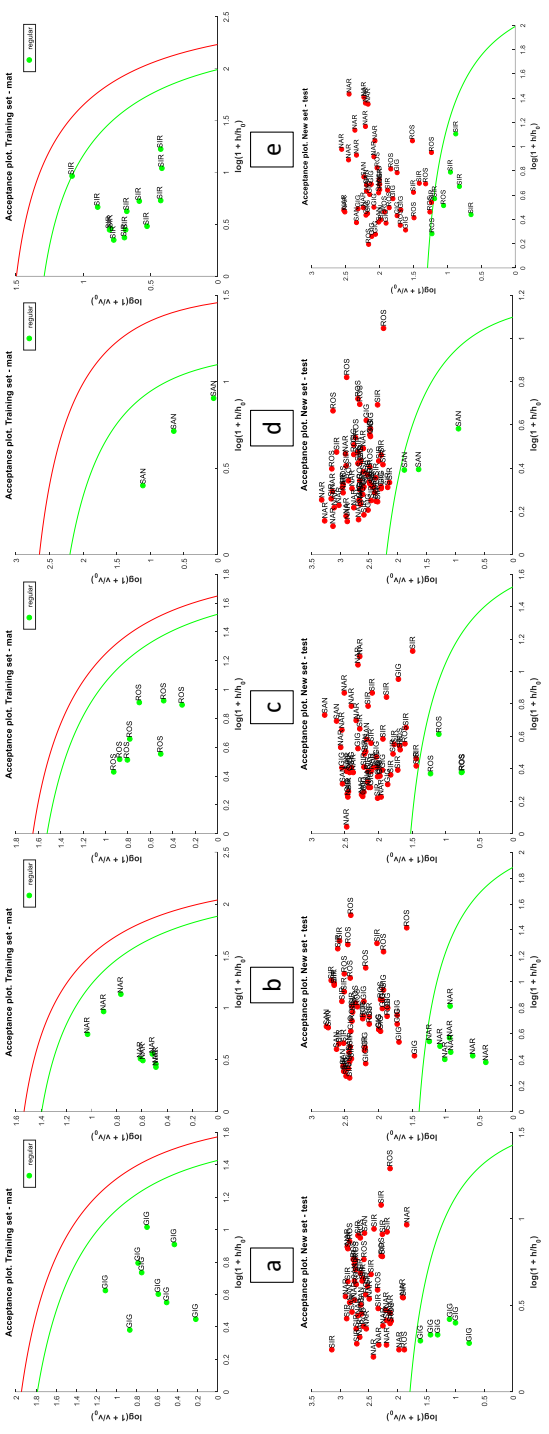


Figure 3. Application of DD-SIMCA models for each groups authentication. *a, b, c, d* and *e* corresponds to models for authentication of GIGANTE (α=0,01, 3 components), NARANJAL (α=0,01, 2 components), ROSARIO (α=0,01, 3 components), SAN ANTONIO (α=0,05, 2 components) and SIRENA (α=0,017, 3 components). In the upper part, training set with first year samples. In the lower part, validation set for the rest of the samples for both harvesting years.

of different years. Instead of using QC samples to create the correction vector, the sample set was used to normalize between different years.

This normalization is mandatory due to the differences in the instrument between two different periods of time, as well as differences in column performance or other instrumental features. With this normalization, all the features in each sequence were located between 0 and 1, making all features comparable between sample batches, without interferences in the number of samples from each geographical zone, although this number was tried to be balanced with the amount of samples employed in the model. After this normalization, PLS-DA and DD-SIMCA models were created with the samples from the first year with PQN normalization.

The PLS-DA model for RP- with the manually integrated data showed a goodness-of-fit R^2Y = 0.841 and goodness-of-prediction Q^2Y= 0.761, closely similar to that obtained with XCMS integrated data (R^2Y = 0.849 and Q^2Y = 0.689).

The cross validation of the model yielded 39 correctly classified samples from a total of 41 samples included in the model (39/41), showing 95 % of correctly classified samples. When the additional validation samples were introduced in the model (**Figure 1b**), 30/32 were correctly classified while 2/32 were assigned as unknown (two *Gigante* samples not classified in any group), showing a 94 % of samples with the correct origin.

At this point, it is important to have a model available that does not lead to false positives, as it is better to have methods which classify samples as unknowns ("false negatives") than give the samples a false origin. In case of unknown samples, the model showed more than one possibility in the less likely classification and an additional classification model should be required.

Regarding validation of DD-SIMCA models, as they were created taking into account only one class, the rest of classes are always treated as "alien" samples. Thus, DD-SIMCA models are less prone to misclassify any new sample belonging to other potential classes. For this reason, DD-SIMCA models have been validated with additional samples (second season) with first-season samples of different classes also included. As can be seen in **Figure 3c** bottom for validation tests, only two *Rosario* samples were not correctly classified. These samples were misclassified as *Sirena* (false

positives). Regarding *Sirena* samples (**Figure 3e** bottom), two of them were not classified in its model but not misclassified as before. In this sense, 4/73 samples were not correctly classified in the models validation achieving a 94% of samples correctly classified.

Authentication methods such as DD-SIMCA work better against alien samples, as reported in the literature and corroborated in the present work, but PLS-DA has also demonstrated its power to correctly classify samples belonging to classes included in the model with similar success rate. Nevertheless, in real-world situations where food-fraud will be based usually on low-quality products not considered initially, DD-SIMCA works better for authentication purposes than PLS-DA regarding its capabilities to deal with these "alien" samples.

Elucidation of biomarkers and biological meaning

Elucidation of the different markers highlighted was made from accurate-mass data obtained in previously acquired MSE high energy function and additional MS/MS experiments, which were carried out at different collision energies (10, 20, 30 and 40 eV). Results are shown in **Table 1**.

As an illustrative example of the elucidation process using HRMS(/MS) accurate-mass data combined with chemical and mass spectral databases, **Figure 4** summarizes the results for marker M353T225. Its accurate mass (*m/z* 353.0858) was retrieved from XCMS table (feature list). Elemental composition calculated using Masslynx software in LE mass spectra function and the i-FIT method resulted on "C16H18O9" as the molecular formula (mass error -2.3 ppm). MS/MS spectra were used to identify product ions coming from precursor ion *m/z* 353 as *m/z* 191.0549, 179.0340, 173.0451, 135.0440, 127.0392, 111.0433, 93.0334 and 85.0280 (among others). The product ions were compared with *Metlin, HMDB and FoodDB* spectral MS/MS library as well as with *Metfrag* (in-silico fragmentation for computer assisted identification of metabolite mass spectra) in order to justify the fragments obtained. After observing results in *Metlin,* marker M353T225 was tentatively elucidated as a chlorogenic acid (CQA). Then, based on the work of Parveen et. al. (*Parveen, Threadgill, Hauck, Donnison, & Winters, 2011*) and the different product ions in each isomer observed at the HE function from MSE (**Figure 5**), it was tentatively elucidated as cis 5-caffeoylquinic acid (5-CQAcis). 5-CQA is a compound widely recognized in the cup coffee chemistry to contribute to coffee quality and health

benefits (*Zanin, Corso, Kitzberger, Scholz, & Benassi, 2016*) and it has been also employed to compare the quality of Colombian Coffee with other countries (*Wei et al., 2012*). Chlorogenic analogs are a

Table 1. Elucidation information for markers used in the PLS-DA analysis

VIP	Compound	Feature name	Elemental composition	RT (min)	Experimental m/z	Theoretical m/z	Error (mDa/ppm)	Ion/Adduct detected
1	---------	M147T64	$C_5H_7O_5$	1.07	147.0290	147.0293	-0.3/-2.0	$[M-H]^-$
2	1-O-Sinapoylglucose	M385T106	$C_{17}H_{22}O_{10}$	1.77	385.1137	385.1135	0.0/0.0	$[M-H]^-$
3	3-Hydroxysuberic acid	M189T301	$C_8H_{14}O_5$	5.02	189.0760	189.0763	-0.3/-1.5	$[M-H]^-$
4	---------	M533T56	$C_{19}H_{33}O_7$	0.94	533.1663	533.1659	+0.4/+0.8	$[M-H]^-$
5	---------	M424T303	$C_{19}H_{23}NO_{10}$	5.04	424.1241	424.1244	-0.3/-0.7	$[M-H]^-$
6	---------	M369T409	---------	6.82	369.1180	---------	---------	---------
7	N-Acetyl-L-Phenylalanine	M206T292	$C_{11}H_{13}NO_3$	4.87	206.0814	206.0817	-0.4/-1.9	$[M-H]^-$
8	5-Caffeoyl-Methylquinic acid (5-FQA$_{trans}$)	M735T249	$C_{34}H_{39}O_{18}$	4.15	735.2147	735.2136	+1.1/+1.5	$M{\cdot}[M-H]^-$
9	Caffeoyl alcohol	M165T278	$C_9H_{10}O_3$	4.63	165.0549	165.0552	-0.4/-2.4	$[M-H]^-$
10	5-Caffeoylquinic acid (5-CQA$_{cis}$)	M353T225	$C_{16}H_{18}O_9$	3.75	353.0866	353.0873	-0.8/-2.3	$[M-H]^-$
11	---------	M467T497	---------	8.28	467.2118	---------	---------	---------
12	5-Caffeoyl-Methylquinic acid (5-FQA$_{cis}$)	M367T296	$C_{17}H_{20}O_9$	4.94	367.1025	367.1029	-0.4/-1.1	$[M-H]^-$
13	Palmitic acid	M255T980	$C_{16}H_{31}O_2$	16.34	255.2319	255.2324	-0.5/-2.0	$[M-H]^-$

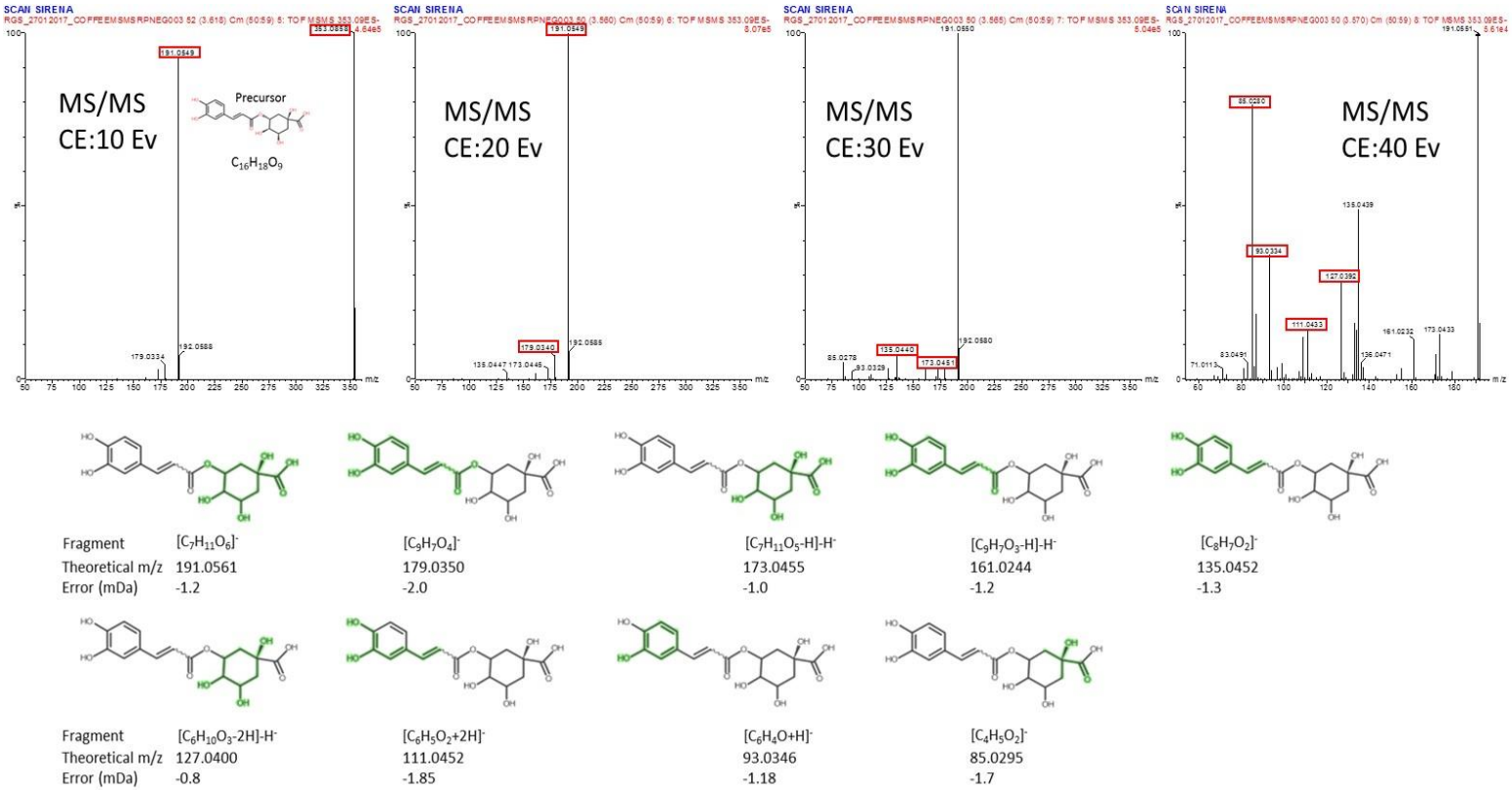


Figure 4. MSMS spectra for marker M353T225 at 10, 20, 30 and 40 eV and fragment explanation using in-silico fragmentation (Metfrag)

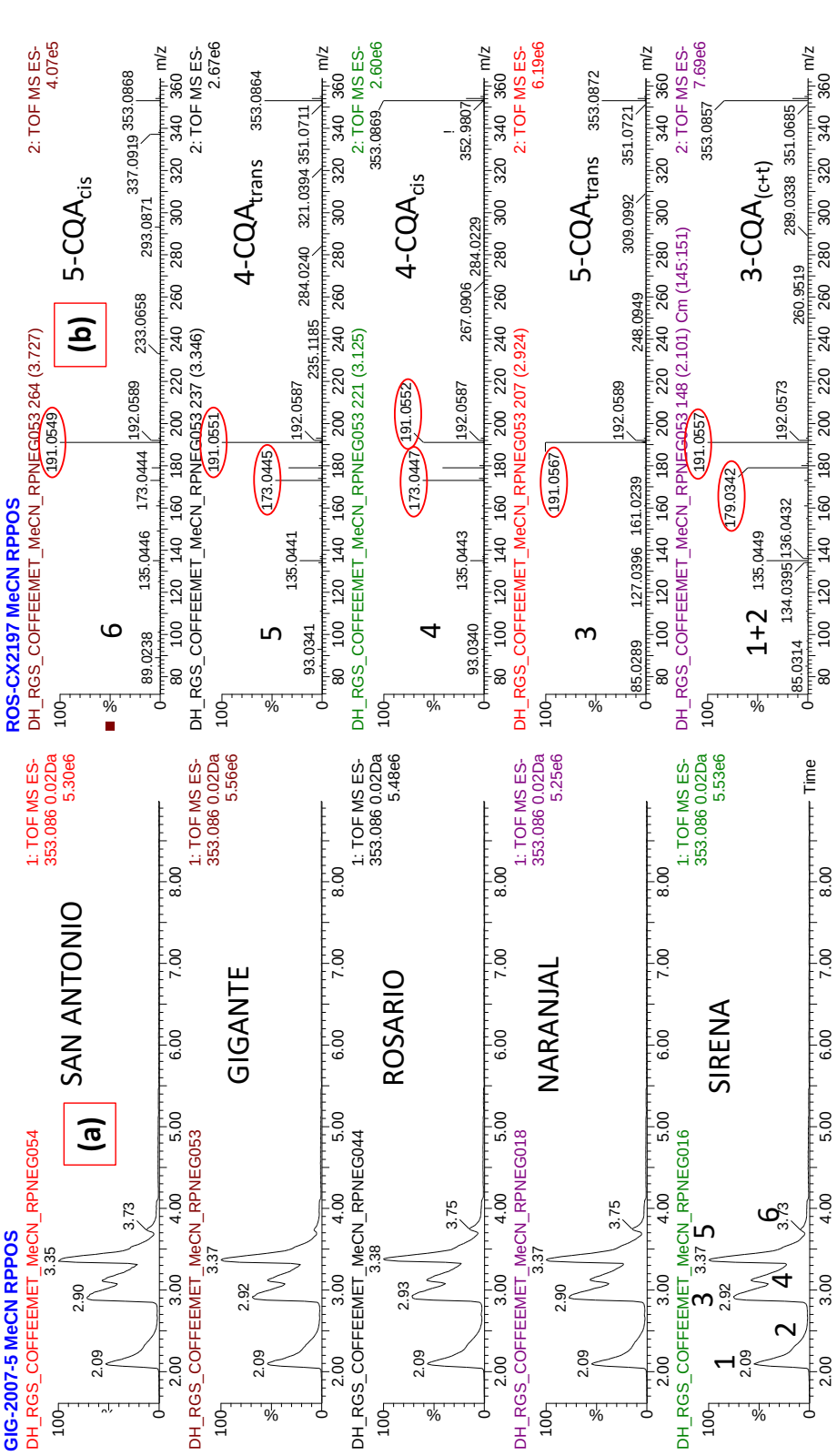


Figure 5: Elucidation of marker #10 M353T225: (a) XIC's in each sample zone; (b) spectrum from each peak. Specific product ions for each isomer are marked in red.

family of esters from quinic acid (QA) with endogenous acids such as caffeic acid (rendering CQA analogue), ferulic acid (FQA) and p-coumaric acid (Narita & Inouye, 2014), and reported in plants like apple, blackberry, broccoli, potato, coffee, pear, among others. Ester bonds in QA are usually formed in positions 3, 4 or 5, resulting in the known 3-CQA, 4-CQA, 5-CQA, 3-FQA, 4-FQA and 5-FQA esters of caffeic and ferulic acids, respectively.

The markers M735T249 and M367T296 were also tentatively identified as feruloylquinic acid isomers (M735T249 as 5-FQAtrans, and M367T296 as 5-FQAcis (see **Figure S5**). They were useful for green coffee separation based on the geographical origin. 5-FQAtrans was highlighted as a dimer M·(M-H)- probably due to the high concentration compared to 5-FQAcis, as can be observed in **Figure S6**.

Although different CQA isomers were observed, just one was remarkable for the groups separation. However, due to lack of CQA standards (3-CQA, 4-CQA and 5-CQA) the highest level for confirmation was not achieved (*Schymanski et al., 2014*) and they were just tentatively elucidated. The same situation occurred for FQA, where 5-FQA cis/trans isomers were just tentatively elucidated.

From the 13 markers used in the PLS-DA model, 8 were tentatively elucidated using accurate mass, HE function in MSE and MS/MS experiments. Unequivocal confirmation would require the use of reference standards injected under the same conditions for retention time and fragmentation evaluation. Although several markers were related with health and quality benefits, from an authentication point of view, this fact is somewhat irrelevant and even harmful compounds could be good discriminant markers.

Once these markers were established, the next step would be to develop and apply a quantitative method for these compounds, using preferably LC-MS/MS with triple quadrupole, a task that does not seem complicated once the markers are identified in our untargeted metabolomics approach.

Conclusions

An untargeted metabolomics approach based on high resolution mass spectrometry coupled to UHPLC has been applied to Colombian green coffee samples for construction of a discrimination model based on the fraction extracted with H_2O:ACN. PLS-DA and DD-SIMCA statistical models were used to classify green coffee samples harvested at different geographical zones inside the Colombian territory. 13 biomarkers were identified (8 of them tentatively elucidated) based on accurate mass measurements resulting from HE (MSE) and MS/MS experiments. The validation-prediction ability of the models was evaluated, showing more than 94% of correctly classified samples in both cases. However, DD-SIMCA demonstrated better capabilities to deal with new classes not used during model building.

The results of our work reveal that UHPLC-(Q)Tof MS-based metabolomics is a suitable approach for green coffee origin discrimination. Once the markers have been established by the untargeted metabolomics approach, it will be possible to apply a quantitative method for these compounds, using preferably LC-MS/MS with triple quadrupole, in a targeted way.

Acknowledgments

The authors acknowledge the collaboration and samples provided by Andres Mauricio Villegas from Centro Nacional de Investigaciones de Café (CENICAFE). The financial support of Sustainability Research Fund of the 2016-2017 administration of the University of Antioquia is appreciated. Authors from IUPA, University Jaume I, acknowledge the support from Generalitat Valenciana (Group of Excellence Prometeo II/2017/023) and Universitat Jaume I (UJI-B2016-10).

References

Alonso-Salces, R. M., Serra, F., Remero, F., & Heberger, K. (2009). Botanical and geographical characterization of green coffee (*Coffea arabica and Coffea canephora*): Chemometric evaluation of phenolic and methylxanthine contents. *Journal of Agricultural and Food Chemistry*, 57(10), 4224–4235.

Arana, V. A., Medina, J., Alarcon, R., Moreno, E., Heintz, L., Schäfer, H., & Wist, J. (2015). Coffee's country of origin determined by NMR: the Colombian case. *Food Chemistry*, 175, 500–6.

Arana, V. A., Medina, J., Esseiva, P., Pazos, D., & Wist, J. (2016). Classification of Coffee Beans by GC-C-IRMS, GC-MS, and 1H-NMR. *Journal of Analytical Methods in Chemistry*, 2016.

Ballabio, D., & Consonni, V. (2013). Classification tools in chemistry. Part 1: linear models. PLS-DA. *Analytical Methods*, 5(16), 3790–3798.

Castillo, L. F., Cristancho, M., Isaza, G., Pinzón, A., & Rodríguez, J. M. C. (Eds.). (2014). Advances in Computational Biology (Vol. 232). Cham: Springer International Publishing.

Castro-Puyana, M., & Herrero, M. (2013). Metabolomics approaches based on mass spectrometry for food safety, quality and traceability. *TrAC Trends in Analytical Chemistry*, 52, 74–87.

Chinnici, F., Durán-Guerrero, E., & Riponi, C. (2015). Discrimination of some European vinegars with protected denomination of origin as a function of their amino acids and biogenic amines content. *Journal of the Science of Food and Agriculture*.

Dettmer, K., Aronov, P. A., & Hammock, B. D. (2007). Mass spectrometry-based metabolomics. *Mass Spectrometry Reviews*, 26(1), 51–78.

Di Guida, R., Engel, J., Allwood, J. W., Weber, R. J. M., Jones, M. R., Sommer, U., … Dunn, W. B. (2016). Non-targeted UHPLC-MS metabolomic data processing methods: a comparative investigation of normalisation, missing value imputation, transformation and scaling. *Metabolomics*, 12(5), 1–14.

Díaz, R., Gallart-Ayala, H., Sancho, J. V, Nu??ez, O., Zamora, T., Martins, C. P. B., … Checa, A. (2016). Told through the wine: A liquid chromatography-mass spectrometry interplatform comparison reveals the influence of the global approach on the final annotated metabolites in non-targeted metabolomics. *Journal of Chromatography A*, 1433, 90–97.

Fandos-Herrera, C. (2016). Exploring the mediating role of trust in food products with Protected Designation of Origin. The case of ´Jamón de Teruel´. *Spanish Journal of Agricultural Research*, 14(1), e0102.

Gallart-Ayala, H., Chéreau, S., Dervilly-Pinel, G., & Le Bizec, B. (2015). Potential of mass spectrometry metabolomics for chemical food safety. *Bioanalysis*, 7(1), 133–46.

García L, J. C., Posada-Suárez, H., & Läderach, P. (2014). Recommendations for the regionalizing of coffee cultivation in Colombia: a methodological proposal based on agro-climatic indices. *PloS One*, 9(12), e113510.

Gil-Solsona, R., Raro, M., Sales, C., Lacalle, L., Diaz, R., Ibañez, M., … Hernández, F. J. (2016). Metabolomic approach for Extra virgin olive oil origin discrimination making use of ultra-high performance liquid chromatography - Quadrupole time-of-flight mass spectrometry. *Food Control*, 70, 350–359.

Goodacre, R., Broadhurst, D., Smilde, A. K., Kristal, B. S., Baker, J. D., Beger, R., … Wulfert, F. (2007). Proposed minimum reporting standards for data analysis in metabolomics. *Metabolomics*, 3(3), 231–241.

Jumhawan, U., Putri, S. P., Yusianto, Marwani, E., Bamba, T., & Fukusaki, E. (2013). Selection of discriminant markers for authentication of Asian palm civet coffee (*Kopi Luwak*): a metabolomics approach. *Journal of Agricultural and Food Chemistry*, 61(33), 7994–8001.

Montero-Vargas, J. M., González-González, L. H., Gálvez-Ponce, E., Ramírez-Chávez, E., Molina-Torres, J., Chagolla, A., Winkler, R. (2013). Metabolic phenotyping for the classification of coffee trees and the exploration of selection markers. *Molecular bioSystems*, 9(4), 693–9.

Nagai, D. K., Santini Pigatto, G. A., & Lourenzani, A. E. B. S. (2016). Formas de inovação na agricultura: O caso da denominação de origem protegida na produção de café de cerrado mineiro. *Espacios*, 37(9).

Narita, Y., & Inouye, K. (2014). Chlorogenic Acids from Coffee. Coffee in Health and Disease Prevention. Elsevier Inc.

Oberthür, T., Läderach, P., Posada, H., Fisher, M. J., Samper, L. F., Illera, J., Collet, L., Moreno, E., Alarcón, R., Villegas, A., Usma, H., Perez, C., & Jarvis, A. (2011). Regional relationships between inherent coffee quality and growing environment for denomination of origin labels in Nariño and Cauca, Colombia. *Food Policy*, 36(6), 783–794.

Parveen, I., Threadgill, M. D., Hauck, B., Donnison, I., & Winters, A. (2011). Isolation, identification and quantitation of hydroxycinnamic acid conjugates, potential platform chemicals, in the leaves and stems of *Miscanthus×giganteus* using LC–ESI-MSn. *Phytochemistry*, 72(18), 2376–2384.

Riedl, J., Esslinger, S., & Fauhl-Hassek, C. (2015). Review of validation and reporting of non-targeted fingerprinting approaches for food authentication. *Analytica Chimica Acta*, 885, 17–32.

Risticevic, S., Carasek, E., & Pawliszyn, J. (2008). Headspace solid-phase microextraction–gas chromatographic–time-of-flight mass spectrometric methodology for geographical origin verification of coffee. *Analytica Chimica Acta*, 617(1–2), 72–84.

Rodionova, O. Y., Titova, A. V., & Pomerantsev, A. L. (2016, April 1). Discriminant analysis is an inappropriate method of authentication. *TrAC - Trends in Analytical Chemistry*. Elsevier.

Rodrigues, C., Brunner, M., Steiman, S., Bowen, G. J., Nogueira, J. M. F., Gautz, L., … Máguas, C. (2011). Isotopes as Tracers of the Hawaiian Coffee-Producing Regions. Journal of Agricultural and *Food Chemistry*, 59(18), 10239–10246.

Rubert, J., Lacina, O., Zachariasova, M., & Hajslova, J. (2016). Saffron authentication based on liquid chromatography high resolution tandem mass spectrometry and multivariate data analysis. *Food Chemistry*, 204, 201–9.

Sales, C., Cervera, M. I., Gil, R., Portol??s, T., Pitarch, E., & Beltran, J. (2017). Quality classification of Spanish olive oils by untargeted gas chromatography coupled to hybrid quadrupole-time of flight mass spectrometry with atmospheric pressure chemical ionization and metabolomics-based statistical approach. *Food Chemistry*, 216, 365–373.

Schymanski, E. L., Jeon, J., Gulde, R., Fenner, K., Ru, M., Singer, H. P., & Hollender, J. (2014). Identifying Small Molecules via High Resolution Mass Spectrometry: Communicating Confidence. *Environmental Science & Technology*, 48(4), 2097–2098.

Serra, F., Guillou, C. G., Reniero, F., Ballarin, L., Cantagallo, M. I., Wieser, M., … Vanhaecke, F. (2005). Determination of the geographical origin of green coffee by principal component analysis of carbon, nitrogen and boron stable isotope ratios. *Rapid Communications in Mass Spectrometry*, 19(15), 2111–2115.

Szymańska, E., Saccenti, E., Smilde, A. K., & Westerhuis, J. A. (2012). Double-check: Validation of diagnostic statistics for PLS-DA models in metabolomics studies. *Metabolomics*, 8, 3–16.

Valentin, J. L., & Watling, R. J. (2013). Provenance establishment of coffee using solution ICP-MS and ICP-AES. *Food Chemistry*, 141(1), 98–104.

van den Berg, R. a, Hoefsloot, H. C. J., Westerhuis, J. a, Smilde, A. K., & van der Werf, M. J. (2006). Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics*, 7, 142.

Veselkov, K. A., Vingara, L. K., Masson, P., Robinette, S. L., Want, E., Li, J. V., … Nicholson, J. K. (2011). Optimized preprocessing of ultra-performance liquid chromatography/mass spectrometry urinary metabolic profiles for improved information recovery. *Analytical Chemistry*, 83(15), 5864–5872.

Villarreal, D., Laffargue, A., Posada, H., Bertrand, B., Lashermes, P., & Dussert, S. (2009). Genotypic and environmental effects on coffee (*Coffea arabica L.*) bean fatty acid profile: impact on variety and origin chemometric determination. *Journal of Agricultural and Food Chemistry*, 57(23), 11321–7.

Weckerle, B., Richling, E., Heinrich, S., & Schreier, P. (2002). Origin assessment of green coffee (*Coffea arabica*) by multi-element stable isotope analysis of caffeine. *Analytical and Bioanalytical Chemistry*, 374(5), 886–890.

Wei, F., Furihata, K., Koda, M., Hu, F., Kato, R., Miyakawa, T., & Tanokura, M. (2012). C NMR-Based Metabolomics for the Classification of Green Coffee Beans According to Variety and Origin. *Journal of Agricultural and Food Chemistry*, 60, 10118–10125.

Westerhuis, J. A., Hoefsloot, H. C. J., Smit, S., Vis, D. J., Smilde, A. K., Velzen, E. J. J., Dorsten, F. A. (2008). Assessment of PLSDA cross validation. *Metabolomics*, 4(1), 81–89.

Yi, L., Dong, N., Yun, Y., Deng, B., Ren, D., Liu, S., & Liang, Y. (2016). Chemometric methods in data processing of mass spectrometry-based metabolomics: A review. *Analytica Chimica Acta*, 914, 17–34.

Zanin, R. C., Corso, M. s. P., Kitzberger, C. n. S. G., Scholz, M. B. g. d. S., & Benassi, M. d. T. (2016). Good cup quality roasted coffees show wide variation in chlorogenic acids content. LWT - *Food Science and Technology*, 74, 480–483.

Zontov, Y. V., Rodionova, O. Y., Kucheryavskiy, S. V., & Pomerantsev, A. L. (2017). DD-SIMCA – A MATLAB GUI tool for data driven SIMCA approach. *Chemometrics and Intelligent Laboratory Systems*, 167, 23–28.

Supplementary material

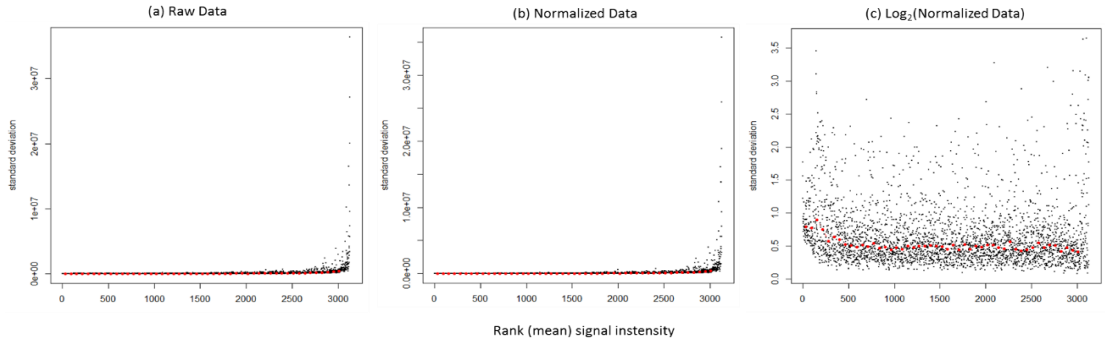


Figure S1. Standard deviation as a function of Rank (mean) signal intensity for samples acquired in negative ionization mode for samples analyzed by RP chromatography: (a) Raw data, (b) Normalized data and (c) Normalized and transformed (log_2) data.

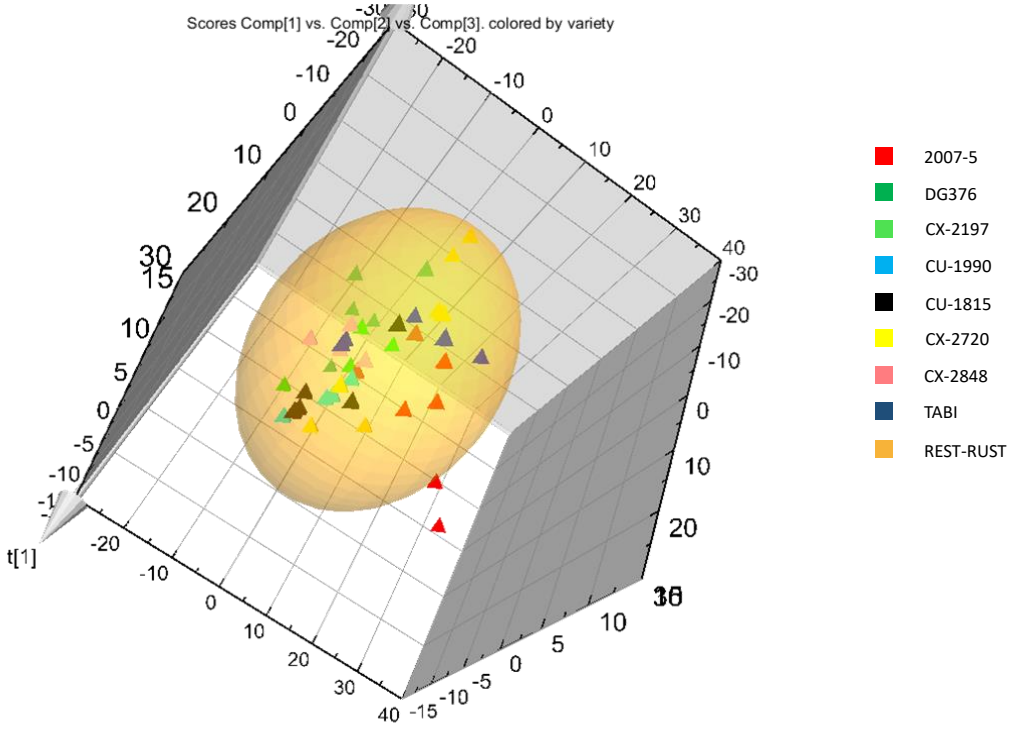


Figure S2. Unsuccessfully PLS-DA model using genotypes.

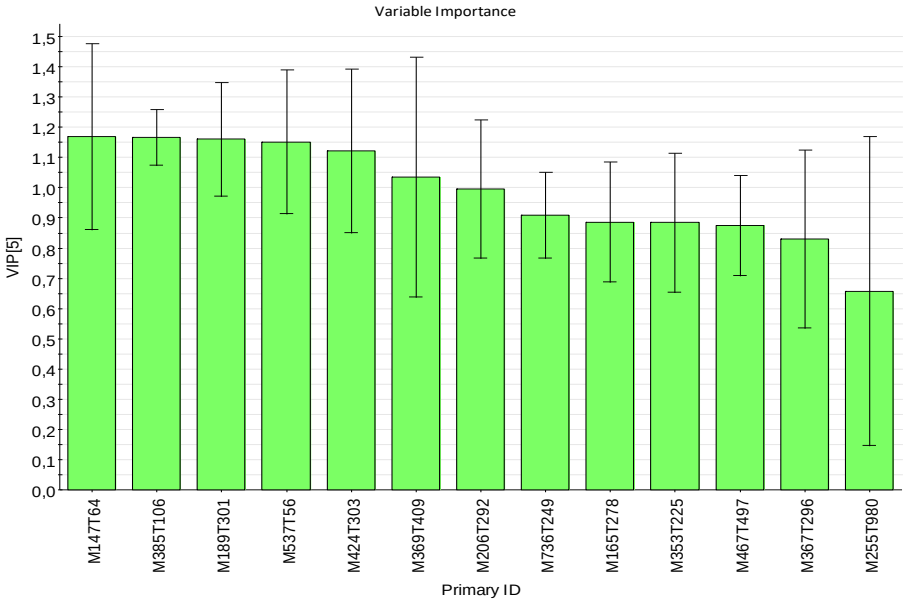


Figure S3. VIP ranking for PLS-DA RP- model

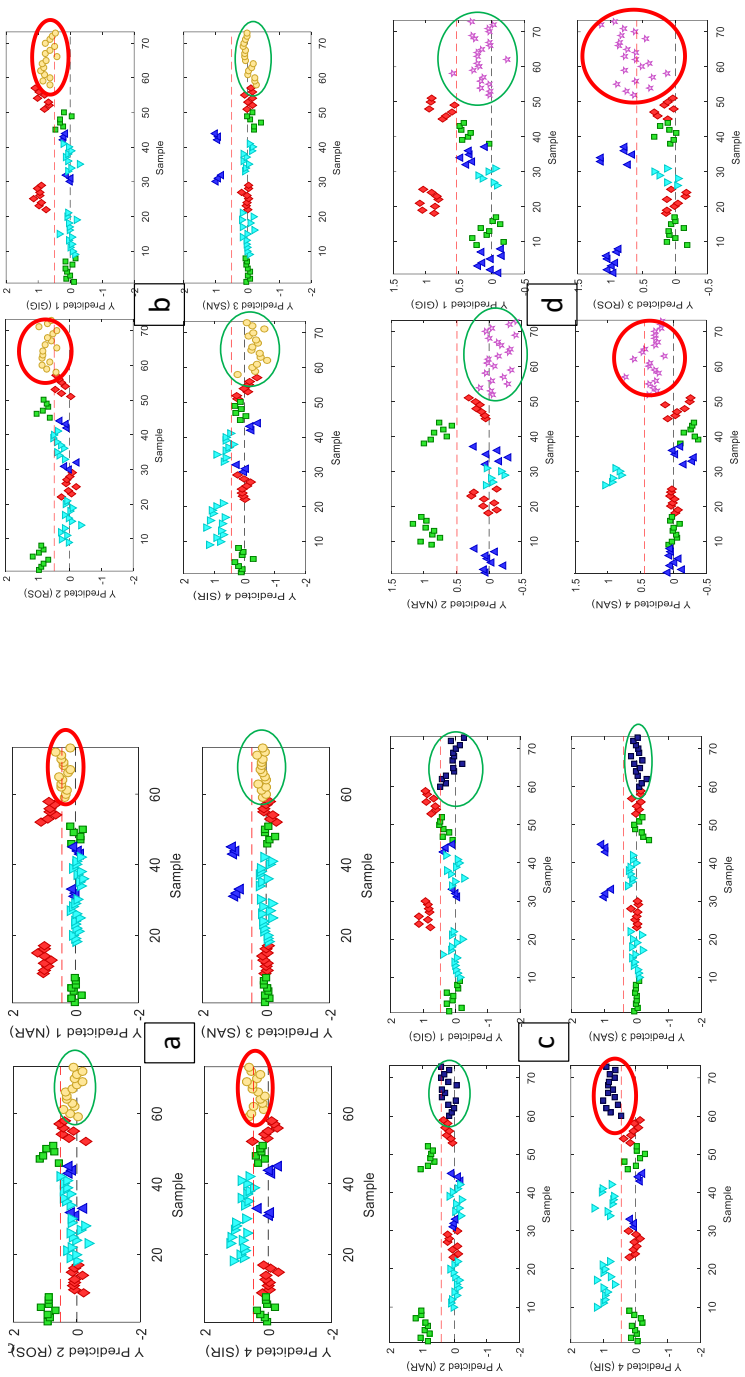


Figure S4. a) PLS-DA score plot without *Gigante* samples. b) PLS-DA score plot without *Naranjal* samples. c) PLS-DA score plot without *Rosario* samples. d) PLS-DA score plot without *Sirena* samples. In all the cases, rounded samples are the class not included in the model. As can be observed, PLS-DA did not work properly (circled in red) for classifying "alien" samples.

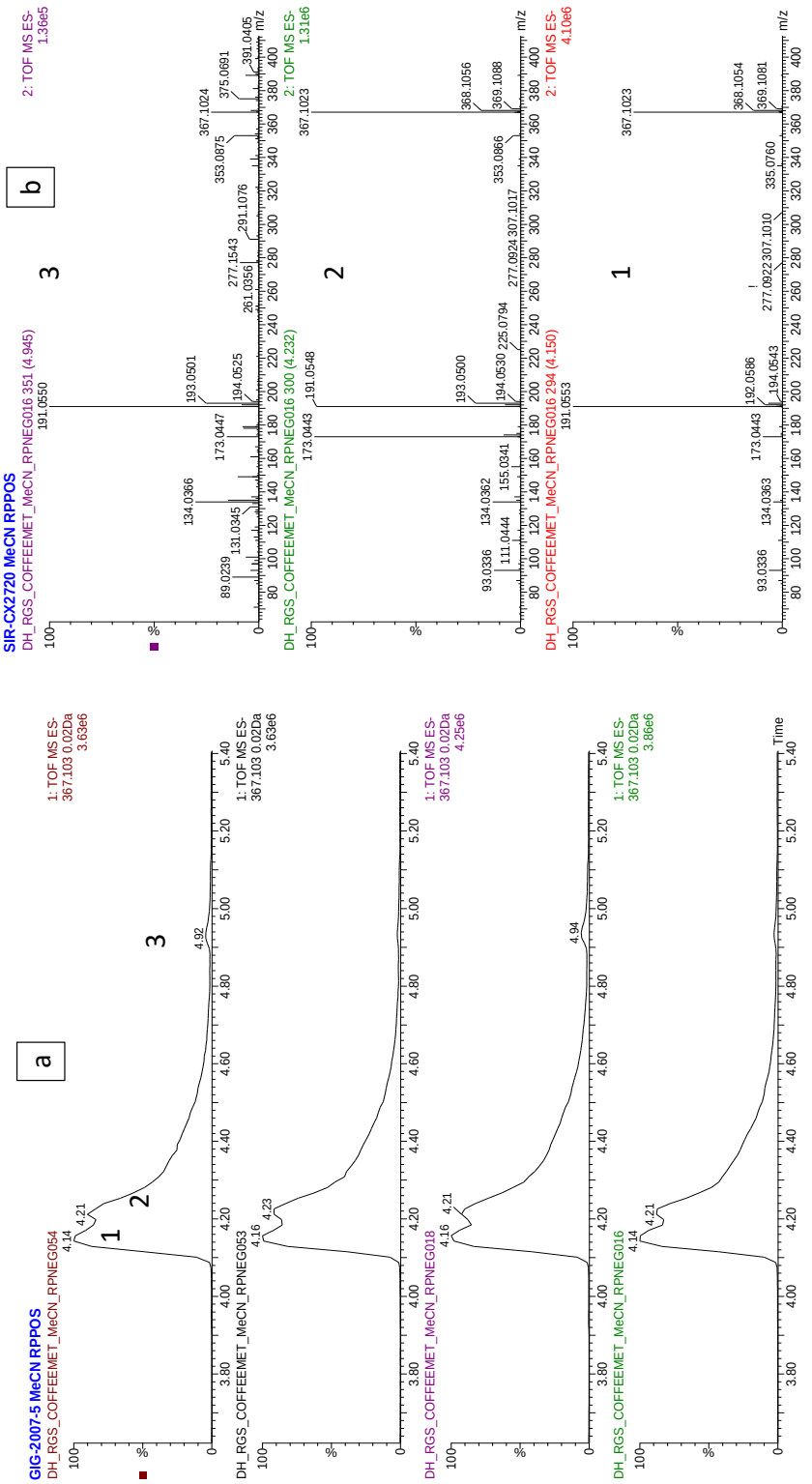


Figure S5. The markers M735T249 (peak 1) and M367T296 (peak 3) tentatively identified as feruloylquinic acid isomers

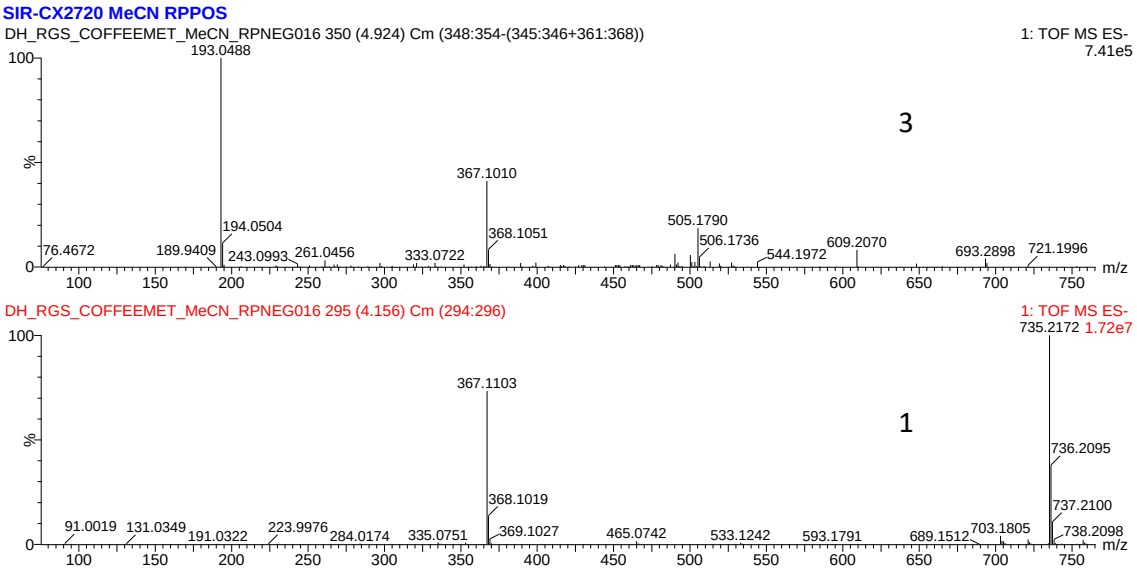


Figure S6. Accurate mass spectra for markers M735T249 (peak 1) and M367T296 (peak 3) tentatively identified as feruloylquinic acid isomers

CAPÍTULO IV

OBTENCIÓN DE BIOMARCADORES DE CONSUMO DE NUEVAS SUSTANCIAS PSICOACTIVAS

IV.1: Artículo científico 6

Analytical and Bioanalytical Chemistry (2018) 410:5107–5112
https://doi.org/10.1007/s00216-018-1182-8

COMMUNICATION

What about the herb? A new metabolomics approach for synthetic cannabinoid drug testing

Lubertus Bijlsma [1] · Rubén Gil-Solsona [1] · Félix Hernández [1] · Juan Vicente Sancho [1]

Received: 23 April 2018 / Revised: 22 May 2018 / Accepted: 4 June 2018 / Published online: 16 June 2018

Abstract
Synthetic cannabinoids (SCs) are consumed as legal alternative to cannabis and often allow passing drug-screening tests. Their rapid transience on the drug scene, combined with their mostly unknown metabolic profiles, creates a scenario with constantly moving analytical targets, making their monitoring and identification challenging. The development of fast screening strategies for SCs, not directly focused on their chemical structure, as an alternative to the commonly applied target acquisition methods, would be highly appreciated in forensic and public health laboratories. An innovative untargeted metabolomics approach, focused on herbal components commonly used for 'spice' products, was applied. Saliva samples of healthy volunteers were collected at pre-dose and after smoking herbal components and analysed by high-resolution mass spectrometry. The data obtained, combined with appropriate statistical analysis, allowed to highlight and elucidate two markers (scopoletin and *N*,*N*-bis(2-hydroxyethyl)dodecylamine), which ratio permitted to differentiate herbal smokers from non-smokers. The proposed strategy will allow discriminating potential positives, on the basis of the analysis of two markers identified in the herbal blends. This work is presented as a step forward in SC drug testing, promoting a smart first-line screening approach, which will allow reducing the number of samples to be further investigated by more sophisticated HRMS methods.

Keywords Spice · Herbal components · Synthetic cannabinoids · Saliva biomarkers · Metabolomics · High-resolution mass spectrometry

Introduction

New synthetic cannabinoids (SCs) are introduced each year into the international 'market' as legal alternative to cannabis [1, 2] . Although they are structurally unrelated, SCs act functionally similar to Δ^9-tetrahydrocannabinol (THC), i.e. the principal active component of cannabis. Branded herbal or 'spice' products, containing SCs, are easily purchased through online vendors and smart shops, where they are sold with misleading information about their effects and safety. Hence, SCs are considered a growing problem and are associated with severe negative effects in consumer's health [3–5].

The impetus of smoking herbal products, adulterated with SCs, is often not only to get intoxicated legally, but it also enables users to pass drug-screening tests. The latter is an important element for any setting where drug abstinence control is obligatory, e.g. specific psychiatry or prison settings, driving liability testing, or employment drug testing procedures [1].

The detection and identification of SCs is an analytical challenge due to their rapid transience on the drug scene [6, 7]. This fact combined with the mostly unknown and extensive metabolic pathways has created a scenario, where a high number and constantly moving analytical targets need to be considered for a comprehensive investigation of SCs consumption. Around 170 SCs have been reported until 2017 to the Early Warning System of the European Monitoring Centre for Drugs and Drug Addiction (EMCDDA), a number that will surely increase in the next years. In addition, analytical reference standards are not always available or it takes much time to purchase them. This has triggered analytical chemists to

Lubertus Bijlsma and Rubén Gil-Solsona contributed equally to this work.

✉ Lubertus Bijlsma
bijlsma@uji.es

✉ Juan Vicente Sancho
sanchoj@uji.es

[1] Research Institute for Pesticides and Water, University Jaume I, Avda Sos Baynat s/n, 12071 Castellón, Spain

What about the herb? A new metabolomics approach for synthetic cannabinoid drug testing

Lubertus Bijlsma[^*], Rubén Gil-Solsona[^], Félix Hernández, Juan Vicente Sancho[*]

Research Institute for Pesticides and Water, University Jaume I, Castellón, Spain

^ These authors contributed equally.

* Authors for correspondence

Abstract

Background and aim Synthetic cannabinoids (SCs) are consumed as legal alternative to cannabis and often allow passing drug-screening tests. Their rapid transience on the drug scene, combined with their mostly unknown metabolic profiles, creates a scenario with constantly moving analytical targets, making their monitoring and identification challenging. The development of fast screening strategies for SCs, not directly focused on their chemical structure, as an alternative to the commonly applied target acquisition methods, would be highly appreciated in forensic and public health laboratories. **Methods** An innovative untargeted metabolomics approach, focused on herbal components commonly used for `spice´ products, was applied. Saliva samples of healthy volunteers were collected at pre-dose and after smoking herbal components and analyzed by high-resolution mass spectrometry. **Results** The data obtained, combined with appropriate statistical analysis, allowed to highlight and elucidate two markers (Scopoletin and N,N-bis(2-hydroxyethyl)dodecylamine), which ratio permitted to differentiate herbal smokers from non-smokers. **Conclusions** The proposed strategy will allow discriminating potential positives, on the basis of the analysis of two markers identified in the herbal blends. This work is presented as a step forward in SC drug testing, promoting a smart first-line screening approach, which will allow reducing the number of samples to be further investigated by more sophisticated HRMS methods.

Introduction

New synthetic cannabinoids (SCs) are introduced each year into the international `market´ as legal alternative to cannabis (EMCDDA, 2009; UNODC, 2017). Although they are structurally unrelated, SCs act functionally similar to $\Delta^9$-tetrahydrocannabinol (THC) *i.e.* the principal active component of cannabis. Branded herbal or `spice´ products, containing SCs, are easily purchased through online vendors and smart shops, where they are sold with misleading information about their effects and safety. Hence, SCs are considered a growing problem and are associated with severe negative effects in consumer's health (Angerer, Jacobi, Franz, Auwärter, & Pietsch, 2017; Seely, Lapoint, Moran, & Fattore, 2012; van Amsterdam, Brunt, & van den Brink, 2015).

The impetus of smoking herbal products, adulterated with SCs, is often not only to get intoxicated legally, but it also enables users to pass drug-screening tests. The latter is an important element for any setting where drug abstinence control is obligatory e.g. specific psychiatry or prison settings, driving liability testing or employment drug testing procedures (EMCDDA, 2009).

The detection and identification of SCs is an analytical challenge due to their rapid transience on the drug scene (Bijlsma et al., 2017; Uchiyama, Kikura-Hanajiri, Ogata, & Goda, 2010). This fact combined with the mostly unknown and extensive metabolic pathways has created a scenario, where a high number and constantly moving analytical targets need to be considered for a comprehensive investigation of SCs consumption. Around 170 SCs have been reported until 2017 to the Early Warning System of the European Monitoring Centre for Drugs and Drug Addiction (EMCDDA), a number that will surely increase in the next years. In addition, analytical reference standards are not always available or it takes much time to purchase them. This has triggered analytical chemists to develop comprehensive screening strategies mostly based on high-resolution mass spectrometry (HRMS), employing suspect and non-targeted approaches (Grabenauer, Krol, Wiley, & Thomas, 2012; Shanks, Dahn, Behonick, & Terrell, 2012). HRMS has shown strong potential to screen for large lists of target compounds, as well as unknown substances, including in a retrospective manner (Ibáñez et al., 2013; Shanks et al., 2012). Its strong potential for identification and elucidation of compounds relies on accurate-mass full-spectrum data, but HRMS-based methods also have limitations in capacity and

sensitivity (Cannaert, Franz, Auwärter, & Stove, 2017), and data processing is very laborious and time consuming, which makes it increasingly challenging and costly when numerous samples need to be analysed.

The development of alternative, fast and generic screening methods, not directly focused on the analyte chemical structure, is an attractive approach that would facilitate the complex analytical scenario and provide fast response on potential SC consumption. Recently, Canneart et al. developed an assay that allowed activity profiling of SCs and their metabolites (Cannaert et al., 2017). This activity-based assay might serve as a first-line screening tool of urine, complementing conventional targeted and untargeted analytical methods. This strategy does not allow the direct identification of SCs, but is neither limited to a given list of compounds, oppositely to the target approaches, which may lead to reporting false negatives when the compound (parent drug or metabolite) is not included in the candidate list.

In this work, we apply an innovative untargeted metabolomics approach for the rapid monitoring of SCs consumption. The strategy applied is focused on the main natural herbal components used for `spice´ products. Thus, instead of screening for target SCs or unknown substances, identified markers of herbs can be monitored to flag suspect samples, since these herbs are commonly used as the herbal base for the active synthetic chemical ingredients (EMCDDA, 2015). We present the proof-of-principle of this original strategy directed towards saliva samples instead of urine. The collection of a saliva samples is fast, easy and non-invasive, it is also less prone to fraud with respect to urine samples, especially essential in obligatory drug control settings.

Material & methods

Chemicals and reagents

HPLC-grade water was obtained by purifying demineralized water in a Mili-Q plus system from Millipore (Bedford, MA, USA). HPLC-grade acetonitrile (ACN), methanol (MeOH), and ammonium acetate (NH_4Ac) were acquired from Scharlab S.L. (Barcelona, Spain). Leucine-enkephalin, formic acid (HCOOH, 98 - 100 %), Scopoletin (99% purity) and N,N-bis(2-hydroxyethyl)dodecylamine (99% purity) were purchased from Sigma-Aldrich (Augsburg, Germany).

Six natural herbal components *Canavalia maritima, Leonurus sibiricus, Althaea officinalis, Turnera diffusa, Verbascum thapsus, Calendula officinalis* were purchased from Worldherbals (Vlaardingen, the Netherlands). Tobacco from different trademarks (Domingo, Fortuna and Camel) was purchased from a local tobacconist.

Sample collection and treatment

Hand-rolled cigarettes containing 0.5 g of tobacco or mixtures of 0.25 g herb with 0.25 g of tobacco (50:50, w/w), were prepared. Each mixture contained tobacco and solely one herbal component. Three healthy volunteers smoked the tobacco cigarette and six herbal mixtures, leaving three days between experiments. Saliva samples were collected in an Eppendorf tube before (at pre-dose, t = 0) and after smoking (t = 30 minutes), and immediately stored in the dark at −20 °C until analyses (within 1 week). The volunteers were informed, and were involved in the study protocol design, giving their consent.

Prior to analyses, saliva samples were thawed, and 0.5 mL sample was vortexed for 1 min. Subsequently, 1 mL of ACN was added, vortexed for 30 sec, sonicated for 1 min and centrifuged at 10.000 g for 10 min. Supernatant was led to dryness under vacuum and reconstituted in 50 µL of H_2O:ACN (90:10 v/v)(Malkar et al., 2013). Quality control samples (QCs), consisting in a mix of all saliva extracts (blanks and smoked), as well as the sample extracts (10µL) were injected directly into the UHPLC-QTOF MS system.

Instrumentation

A Waters Acquity UPLC system (Waters Corp., Milford, MA, USA) was interfaced to a hybrid quadrupole- orthogonal acceleration- ToF mass spectrometer (Xevo G2 QTof, Waters Corp., Manchester, UK) using a Z-spray electrospray ionization (ESI) interface operating in both positive and negative ionization modes. A capillary voltage of 0.7 kV for ESI+ and 1.5 kV for ESI- mode, and a cone voltage of 20 V were used. MS^E data were acquired over the range m/z 50–1200. MS/MS experiments were performed applying a collision energy of 10, 20 and 30 eV.

The chromatographic separation was performed using a CORTECS® C_{18} fused core column (2.7 µm particle size, 2.1x100 mm) at a flowrate of 0.3 mL/min. The mobile phases used were A - MilliQ water and B - MeOH (both with 0.01% HCOOH). The total run time was 18 min.

Further details on instrument operating conditions both chromatographic and spectrometric can be found elsewhere (Ibáñez et al., 2013).

Data processing

Data (*.raw) were converted to a machine independent format (*.cdf) using the DataBridge application within MassLynx™ (Waters Corp., Milford, MA, USA). The converted data was then processed using XCMS free R package. Peak picking was performed with *centWave*, an algorithm for feature selection, which integrates area, considering a peak width between 4-20 sec., ≥ 3 scans with more than 1000 counts, *s/n* ratio of 10, and a mass error of ≤ 15 ppm. Peaks were grouped in single features using the *retcor()* function based on their retention time and mass error (an initial variation < 15 seconds and < 10 ppm, between samples along the batch was considered as the same feature). Features were labelled as M*XXX*T*YYY* where *XXX* corresponds to the nominal Mass of peak *XXX* while *YYY* matches the retention Time in seconds. Samples were $\log_2$ transformed to reduce heteroscedasticity. Pareto scaling was applied before importing the data for statistical analysis.

Statistical analysis

Multivariate analysis was carried out using EZinfo 2.0 (Umetrics – Sartoris Stedim Biotech, Malmö, Sweden). First, Principal Component Analysis (PCA) was applied to ensure that the data was correctly acquired in both positive and negative ionization mode. Next, all data were merged in a single table for further statistical analysis. Partial Least Squares Discriminant Analysis (PLS-DA) was carried out to ensure that samples can be differentiated. Orthogonal Partial Least Squares – Discriminant Analysis (OPLS-DA) was performed in order to extract a small group of markers to differentiate between herb and tobacco.

Results and discussion

Six natural herbal components commonly used as the herbal base for 'spice' products were selected (EMCDDA, 2009; Ogata, Uchiyama, Kikura-Hanajiri, & Goda, 2013). The SCs are typically blended with or sprayed onto these herbs (EMCDDA, 2015).

Hand-rolled herb cigarettes were prepared as a mixture of tobacco and herb (50:50, w/w). Although differences between "only herb and only tobacco" saliva samples would possibly be more pronounced, tobacco could mask the markers for herbs, as cannabis (and very likely SCs as well) is often mixed with tobacco or consumed in parallel (WHO, 2016). Thus, a strategy focused on discriminating between saliva after smoking tobacco and herb-tobacco mixtures would allow a first screening of potential positive samples containing SCs.

Saliva samples were collected using the same material of the same batch to avoid misinterpretations on assignment of potential markers. During experimental days, the volunteers did not have a strict diet. Since all samples collected from the three volunteers were processed at the same time, differences in diet were compensated and the selected markers could therefore be linked to herb and/or tobacco smoking.

Metabolomics enables to discover and highlight differences between subject groups based on the powerful analytical capabilities of HRMS and bioinformatics (Chekmeneva et al., 2017; Patti, Yanes, & Siuzdak, 2012; Raro et al., 2015). Hence, in this study, an untargeted metabolomics approach has been applied as a data-mining tool, to distinguish between tobacco and herb-tobacco mixtures administration. HRMS data was processed and grouped by multiple peak picking functions and statistical tools, considering retention time and mass error. PCA analysis allowed to eliminate possible outliers (Hotelling's T^2 Ellipse (0.95)) and control the possible instrumental drift along the time. The latter could be done by injecting QCs in the initial part of the batch and after every 10 samples. The QCs (n= 6) were grouped in the center of the PCA plot, indicating the correct acquisition of the data. Further statistical analysis by means of PLS-DA modelling showed differences in saliva samples at pre-dose (blank) and after smoking herbs or tobacco (72 % of variance for the first 2 components) (**Figure 1A**). An S-Plot from OPLS-DA, representing all features, permitted to highlight the best two markers (**Figure 1B**).

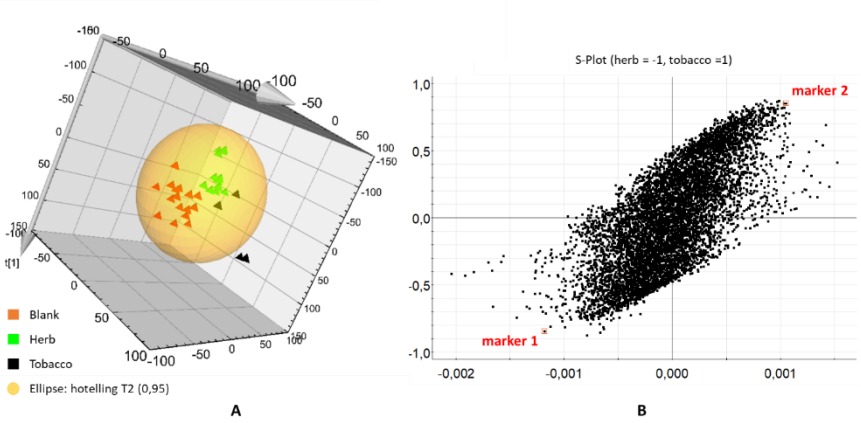


Figure 1: A) PLS-DA 3D score plot for the first three components. B) S-Plot from OPLS-DA where all features are represented

The ratio between the peak area of these two features was evaluated achieving the discrimination between the blank and/or tobacco saliva samples versus the six different herb samples (**Figure 2**). A threshold value of 0.02 could be set (figure 2, insert), meaning that samples with a ratio > 0.02 can be flagged as suspects.

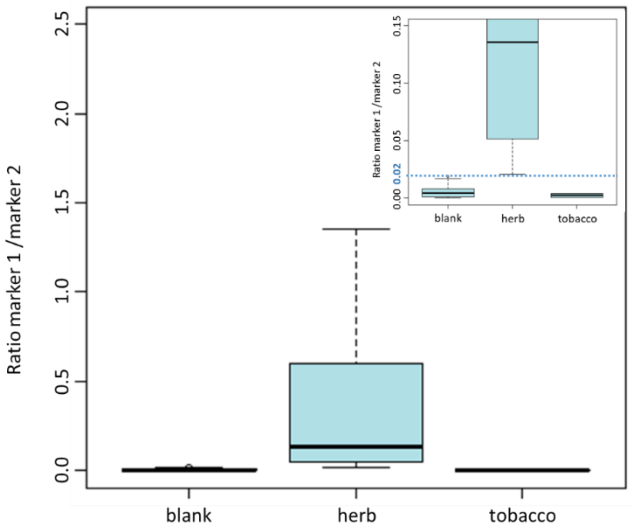


Figure 2: Boxplot representing ratios of the peak areas between marker 1 and marker 2 in the saliva samples *i.e.* blank and after tobacco and herb smoking (n= 24, 6, 18, respectively). Insert; zoom at ratios 0 to 0.15

The elemental compositions of the two most significant markers, retrieved from OPLS-DA, were calculated based on their accurate masses. Subsequently, tandem mass spectrometry (MS/MS) QTOF experiments were performed in order to obtain accurate-mass product-ion spectra. This information was used to search for plausible structures in available mass spectra libraries (Metlin and Massbank), chemical databases (Chemspider), and in-silico fragmentation web sources (MetFrag). **Figure 3** shows the accurate-mass product-ion spectra acquired at a collision energy of 30 eV. The accurate mass of the protonated molecules (mass error < 3 ppm) and of the fragment ions allowed the tentative identification of Scopoletin (marker 1; figure 3 top) and N,N-Bis(2-hydroxyethyl)dodecylamine (marker 2; figure 3 bottom), which were subsequently confirmed by the acquisition of the reference standards.

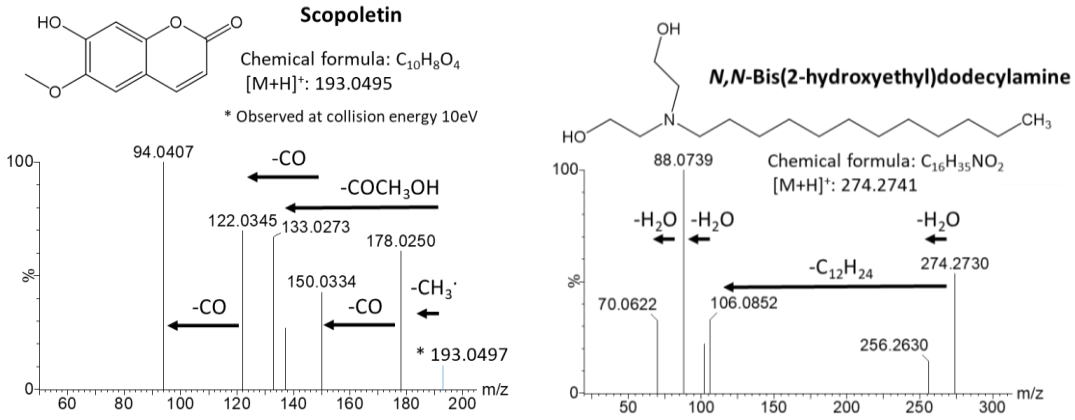


Figure 3: Identification by QTOF operating in MS/MS acquisition mode of Scopoletin (top) and *N,N*-Bis(2-hydroxyethyl)dodecylamine (bottom) in saliva samples. (Inserts: structure, chemical formula and exact mass of the protonated molecule)

Conclusions and future perspectives

In this work, we present an innovative analytical strategy to obtain fast response on potential SC consumption. The application of an untargeted metabolic approach, based on the screening of saliva samples after consumption of tobacco and herbal components commonly used for spicing with SCs, has revealed two biomarkers (scopoletin and N,N-Bis(2-hydroxyethyl)dodecylamine). The ratio between these two biomarkers in all six herbs investigated permitted to discriminate between "tobacco and herb-tobacco mixture" samples. By monitoring their ratio in saliva samples, a value above 0.02 reveals the smoking of herbs, and therefore suggesting, indirectly, the potential consumption of SCs. In this way, samples can be pre-selected for further profound HRMS investigation in order to identify the consumed SCs.

We provide the proof-of-principle for this new approach, which can be considered as an important step forward towards a more generic SC drug test, not directly based on their structures, but on the herbal markers. However, in order to reach that goal future research is needed. This should be focused on the following issues: (i) to understand the significance of the selected biomarkers (already confirmed by their reference standards) and their link to the composition of the herbs investigated; (ii) full validation of the approach suggested, considering more volunteers and

time/data points in order to evaluate more accurately the overall smoke/blood/saliva distribution of the markers; (iii) evaluation of the ratio for the identified markers after smoking other herbs also used as herbal base (i.e. *Nymphaea alba, Scutellaria lateriflora, Zornia latifolia, Nelumbo nucifera, Trifolium pratense, Leonotis leonurus, Astragallus root, Lamiaceae herbs* and *Rosa canina* (EMCDDA, 2015)); (iv) development of target analytical methods for the identified biomarkers in saliva, e.g. based on LC-MS/MS with triple quadrupole.

The strategy suggested is the target analysis (e.g. by LC-MS/MS QqQ) of identified herb biomarkers i.e. indirect indicators of SC consumption, and the subsequent monitoring of their peak area ratio. This approach has two major advantages: i) only positive/suspect samples need further investigation to identify the SCs smoked, which is more rapid and cost-effective since time-consuming analysis and data processing of all samples is avoided. ii) fraudulence in drug-screening tests by SCs consumers becomes more difficult, because monitoring is not based on the chemical structures of individual target SCs, but on the herbs markers. Therefore, it bypasses the fast-changing nature of the SCs market, where the current applications of target analysis might lead to reporting false negatives. In addition, this untargeted metabolomics approach could be applied to other matrices such as breath, hair and urine (Li et al., 2015; Want et al., 2010) for a similar purpose. This might allow the detection of SC consumption over a longer period of time, opening new possibilities for the laboratories to face the complex issue of monitoring the SCs market.

Acknowledgments

Lubertus Bijlsma acknowledges NPS-Euronet (HOME/2014/JDRUG/AG/DRUG/7086), co-funded by the European Union, for his post-doctoral fellowship. This publication reflects the views only of the authors, and the European Commission cannot be held responsible for any use which may be made of the information contained therein. The authors acknowledge the financial support of Generalitat Valenciana (Prometeo II 2014/023) and of the Spanish Ministry of Economy and Competitiveness (Project ref CTQ2015-65603). The authors would like to thank CSM, RGS and RBvL for their collaboration in providing saliva samples.

References

Angerer, V., Jacobi, S., Franz, F., Auwärter, V., & Pietsch, J. (2017). Three fatalities associated with the synthetic cannabinoids 5F-ADB, 5F-PB-22, and AB-CHMINACA. *Forensic Science International*, *281*, e9–e15.

Bijlsma, L., Ibáñez, M., Miserez, B., Ma, S. T. F., Shine, T., Ramsey, J., & Hernández, F. (2017). Mass spectrometric identification and structural analysis of the third-generation synthetic cannabinoids on the UK market since the 2013 legislative ban. *Forensic Toxicology*, *35*(2), 376–388.

Cannaert, A., Franz, F., Auwärter, V., & Stove, C. P. (2017). Activity-Based Detection of Consumption of Synthetic Cannabinoids in Authentic Urine Samples Using a Stable Cannabinoid Reporter System. *Analytical Chemistry*, *89*(17), 9527–9536.

Chekmeneva, E., Dos Santos Correia, G., Chan, Q., Wijeyesekera, A., Tin, A., Young, J. H., … Holmes, E. (2017). Optimization and Application of Direct Infusion Nanoelectrospray HRMS Method for Large-Scale Urinary Metabolic Phenotyping in Molecular Epidemiology. *Journal of Proteome Research*, *16*(4), 1646–1658.

EMCDDA. (2009). *Understanding the `Spice´ phenomenon*.

EMCDDA. (2015). *Perspectives on drugs: Synthetic cannabinoids in Europe*.

Grabenauer, M., Krol, W. L., Wiley, J. L., & Thomas, B. F. (2012). Analysis of synthetic cannabinoids using high-resolution mass spectrometry and mass defect filtering: Implications for nontargeted screening of designer drugs. *Analytical Chemistry*, *84*(13), 5574–5581.

Ibáñez, M., Bijlsma, L., Van Nuijs, A. L. N., Sancho, J. V., Haro, G., Covaci, A., & Hernández, F. (2013). Quadrupole-time-of-flight mass spectrometry screening for synthetic cannabinoids in herbal blends. *Journal of Mass Spectrometry*, *48*(6), 685–694.

Li, X., Martinez-Lozano Sinues, P., Dallmann, R., Bregy, L., Hollmén, M., Proulx, S., … Zenobi, R. (2015). Drug Pharmacokinetics Determined by Real-Time Analysis of Mouse Breath. *Angewandte Chemie - International Edition*, *54*(27), 7815–7818.

Malkar, A., Devenport, N. A., Martin, H. J., Patel, P., Turner, M. A., Watson, P., … Creaser, C. S. (2013). Metabolic profiling of human saliva before and after induced physiological stress by ultra-high performance liquid chromatography–ion mobility–mass spectrometry. *Metabolomics*, *9*(6), 1192–1201.

Ogata, J., Uchiyama, N., Kikura-Hanajiri, R., & Goda, Y. (2013). DNA sequence analyses of blended herbal products including synthetic cannabinoids as designer drugs. *Forensic Science International*, *227*(1–3), 33–41.

Patti, G. J., Yanes, O., & Siuzdak, G. (2012). Metabolomics: the apogee of the omics trilogy. *Nature Reviews Molecular Cell Biology*, *13*, 263–269.

Raro, M., Ibáñez, M., Gil, R., Fabregat, A., Tudela, E., Deventer, K., … Pozo, Ó. J. (2015). Untargeted Metabolomics in Doping Control: Detection of New Markers of Testosterone Misuse by Ultrahigh Performance Liquid Chromatography Coupled to High-Resolution Mass Spectrometry. *Analytical Chemistry*, *87*(16), 8373–8380.

Seely, K. A., Lapoint, J., Moran, J. H., & Fattore, L. (2012). Spice drugs are more than harmless herbal

blends: A review of the pharmacology and toxicology of synthetic cannabinoids. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, *39*(2), 234–243.

Shanks, K. G., Dahn, T., Behonick, G., & Terrell, A. (2012). Analysis of first and second generation legal highs for synthetic cannabinoids and synthetic stimulants by ultra-performance liquid chromatography and time of flight mass spectrometry. *Journal of Analytical Toxicology*, *36*(6), 360–371.

Uchiyama, N., Kikura-Hanajiri, R., Ogata, J., & Goda, Y. (2010). Chemical analysis of synthetic cannabinoids as designer drugs in herbal products. *Forensic Science International*, *198*(1–3), 31–38.

UNODC. (2017). Market analysis of synthetic drugs: Amphetamine-type stimulants, new psychoactive substances. In *World Drug Report 2017* (p. 60).

van Amsterdam, J., Brunt, T., & van den Brink, W. (2015). The adverse health effects of synthetic cannabinoids with emphasis on psychosis-like effects. *Journal of Psychopharmacology*, *29*(3), 254–263.

Want, E. J., Wilson, I. D., Gika, H., Theodoridis, G., Plumb, R. S., Shockcor, J., … Nicholson, J. K. (2010). Global metabolic profiling procedures for urine using UPLC-MS. *Nature Protocols*, *5*(6), 1005–1018.

WHO. (2016). *The health and social effects of nonmedical cannabis use*. *World Health Organization*. Geneva, Switzerland.

IV.2: Artículo científico 7

How ion mobility mass spectrometry adds an extra dimension to metabolomics: finding pyrolytic exposure compounds of synthetic cannabinoids consumption as a case study

Rubén Gil-Solsona[a#], Juan V. Sancho[a#], Anne-Laure Gassner[b], Céline Weyerman[b], Félix Hernández[a], Pierre Esseiva[b], Olivier Delémont[b], Lubertus Bijlsma[a*]

[a] Research Institute for Pesticides and Water (IUPA). Avda. Sos Baynat, s/n. University Jaume I, 12071 Castellón, Spain.

[b] Ecole des Sciences Criminelles, Université de Lausanne, 1015 Lausanne, Switzerland

Abstract

An indirect synthetic cannabinoids control analyzing the herbal components used as base for spice products by means of untargeted metabolomics has been recently proposed. The smoke produced after the herbs combustion has been selected as an easy and readily matrix to work, describing the exposome of spice consumers. The smoke components were trapped in carbon cartridges, desorbed and analysed by liquid chromatography coupled to quadrupole time-of-flight mass analyzer using different separation mechanism (reversed phase and HILIC) and acquiring in both positive and negative mode to widen the polarity scope. Furthermore, the system contained an ion-mobility separator, which added another dimension to the data acquired and which is not much explored in metabolomics until now. Orthogonal Partial Least Squares – Discriminant Analysis highlighted two single compounds whose ratio allowed to differentiate between tobacco and herbal products. The two compounds were tentatively identified using the accurate mass data obtained by the QTOF mass analyzer. In order to gain extra confidence to the tentative identification, retention time and collisional cross section values were predicted using artificial neural networks. The results of this work suggest that UHPLC-IMS-QToF is an efficient instrument for untargeted metabolomics, providing four dimension data, which enhances tentative elucidation confidence in any metabolomics experiment.

Introduction

Untargeted metabolomics have arisen as a powerful analytical tool in the last years. Its workflow is based on discovering and highlighting unknown compounds to differentiate between two or more groups by means of statistical analysis. Metabolomics is based on the combination of powerful techniques with bioinformatics and multivariate statistics, initially appearing for studying metabolite levels in the metabolic cascade of biological scenarios (Dettmer, Aronov, & Hammock, 2007). However, it has rapidly extended to other analytical research fields, such as food analysis (Cevallos-cevallos, Etxeberria, Danyluk, & Rodrick, 2009) drug metabolism (Raro et al., 2015), breath (Couto et al., 2017) and environmental samples (Beale et al., 2016). It enables dealing with complex matrices, emphasising low concentrated substances (e.g. metabolites) among lots of compounds in the same sample and finishing in the annotation of the highlighting compounds.

In this unbiased workflow, probably the main "Achilles heel" is the elucidation process. Despite the combination of separation techniques such as liquid- and gas chromatography with powerful high-resolution accurate-mass mass analyzers (HRMS), which has improved selectivity and mainly sensitivity compared to more classical NMR approaches, elucidation of highlighted compounds is the most time-consuming part. Nevertheless, different tools available when acquiring accurate mass data, such as mass spectra databases and in-silico fragmentation tools, can help the elucidation process to assign possible chemical structures to the candidates. In this sense, most of online databases, such as METLIN (insertar cita de METLIN), contains spectra of biological compounds, naturally occurring in animals or plants, making the tentative identifications of this kind of compounds easier, being also possible to purchase their analytical standards for their confirmation. However, these databases are far from complete and therefore candidate compounds are often not present, especial when candidates result from transformation processes (e.g. degradation, combustion, oxidation, etc). In addition, their reference standards are often not available, so one can only rely on tentative identifications based on well-defined criteria (Schymanski et al., 2014).

In order to achieve more identification confidence, the recent introduction of ion-mobility separators in the core of HRMS instruments (D'Atri et al., 2017) have been used to better define compounds (Kaufmann et. al. 2018). It separates ionized molecules by their collisional cross section (CCS), providing an extra separation dimension to retention time (RT) and accurate mass. Furthermore, the introduction of novel prediction tools using artificial neural networks (ANN) for RT (Bade et al., 2015) and/or CCS values (Bijlsma et al., 2017; Zhou, Tu, Xiong, Shen, & Zhu, 2017), provides an extra elucidation power to tentatively identifications. The use of these machine-based prediction tools can reduce the amount of possible candidates drastically. Only a few papers have been published, applying IMS (Astarita, 2009), showing its power to identify for example lipids (Zhang et al., 2017) and homemade explosives (Hagan et al., 2017), but it surely becomes a relevant tool to be explored more in metabolomics in the coming years.

The objective of this work was to demonstrate the power of combining ultra-high performance liquid chromatography (UHPLC) with ion mobility separation (IMS) and HRMS in untargeted metabolomics studies. To this aim, smoke produced in the combustion of tobacco or other herbs was selected. The herbs selected are often used in spice products, hence the smoke describe the exposome of spice consumers. In this context, the study pretend to analyze, highlight and identify relevant markers of herbs after combustion, since potential unknown pyrolytic compounds can be of relevance of e.g. related health effects or may lead to a tool of indirect synthetic cannabinoids (SCs) control (Bijlsma et. al., 2018). In addition and as described in the workflow above, new RT and CCS predictors for the tentatively identified compounds have been used for reducing the number of possible candidates and obtaining extra confidence in annotation process. Despite the advantages demonstrated, these prediction tools are not much applied in this type of research.

Materials & methods

Chemicals and samples

HPLC-grade water was obtained by purifying demineralized water in a Milli-Q plus system from Millipore (Bedford, MA, USA). HPLC-grade acetonitrile (ACN), Dichloromethane (DCM), methanol (MeOH) and ammonium acetate (NH_4Ac) were obtained from Scharlab (Barcelona, Spain). Leucine-enkephalin, formic acid (HCOOH, 98 - 100 %) and quinoline (98% purity) were purchased from Sigma-Aldrich (Augsburg, Germany).

Fourteen herbs mainly smoked in spice products: *Cannavalia Maritima, Nymphaea alba, Scutellaria Lateriflora, Zornia Latifolia, Nelumbo Nucifera, Leonurus Sibiricus, Althaea Officinalis, Turnera Diffusa, Verbascum Thapsus, Trifolium Pratense, Claendula Officinalis, Leonotis Leonurus, Astragallus root* and *Rosa canina* were purchased from Worldherbals (Vlaardingen, The Netherlands). Tobacco from three different trademarks (*Domingo, Fortuna* and *Camel*) were purchased from a local tobacconist's.

Sample preparation and treatment

All the fourteen herbs as well as three tobacco samples (0.5 g of each one) were rolled in cigarettes (by triplicate) and coupled to an SPE cartridge (ENVI-Carb)®, purchased from Sigma-Aldrich, previously conditioned with 6 mL of MeOH and 6 mL of DCM. Cigarettes were lighted under vacuum without cigarette filter. After extracting the smoke through the cartridge, each one were eluted with 6 mL MeOH:DCM (20:80 v/v). Then it was led nearer to dryness under vacuum using a MiVac Duo concentrator (Genevac, United Kingdom) at low temperature (40ºC, 45 min) in order to minimize losses during this step, and reconstituted with 4 mL of MeOH. All the different herbs and tobacco extractions were carried out by triplicate.

Two aliquots of 0.2 mL were mixed with 1.8 mL Milli-Q water (for Reversed Phase (RP) analysis) for the first aliquot and with 1.8 mL ACN (for HILIC analysis) for the second aliquot. Quality Control (QC) samples were also prepared by pooling all the samples in the same quantity, creating an average sample of all the set.

Instrumentation.

A Waters Acquity UPLC system (Waters, Milford, MA, USA) was interfaced to a hybrid Quadrupole- Ion mobility- Time of Flight (Tof) High Resolution Mass Spectrometer (UHPLC-IMS-HRMS, VION QTof, Waters, Manchester, UK) using a Stepwave interface operating in both positive and negative ionization modes with resolution of the ToF MS approximately 25000 at full width half maximum (FWHM).

Instrumental conditions

UHPLC analysis

Two different UHPLC separations were performed in order to cover a wide range of compound polarities in both experiments. RP Liquid Chromatography (Phenomenex Kinetex 2.6µm C_{18} 100Å, 2.1x100 mm fused core column) was used to separate semi-polar compounds while Hydrophobic Interaction Liquid Chromatography (HILIC) (CORTECS® HILIC 2.7 µm, 2.1x100 mm fused core column) for polar compounds analysis. Gradients and conditions are shown in Table 1.

IMS-QTof MS analysis (VION instrument)

Electrospray (ESI) was employed as interphase, for which capillary voltage was set at 0.7 kV for ESI positive and 1.5 kV for ESI negative ionization modes respectively and 25 V were set as cone voltage. Source temperature was set at 130 °C, (N_2) was employed as desolvation gas with a flow of 800L/h and 550 °C. Argon was employed as collision gas (Purity 99.995%, Carbagas, Lausanne, Switzerland). For $HDMS^E$ experiments, two acquisition functions were configured, with different collision energies: Low energy function (LE), selecting 6 eV and high energy function (HE) with a ramp of collision energies from 15 to 40 eV. MS data were acquired over an *m/z* range of 50-1200 Da.

Table 1. Chromatographic conditions for RPLC and HILIC analysis.

	RPLC analysis		HILIC analysis	
Mobile phase	A: 0.1% HCOOH Milli-Q H_2O B: 0.1% HCOOH MeOH		A: 0.1% HCOOH 10 mM NH_4Ac ACN B: 0.1% HCOOH 10 mM NH_4Ac Milli-Q H_2O	
	Time (min)	%B	Time (min)	%B
	0.00	10	0.00	2
	14.00	90	1.50	2
	16.00	90	2.50	15
	16.01	10	6.00	50
			7.50	60
			8.00	2
	Total run time: 18 min		Total run time: 18 min	
	Flow: 0.3 mL/min Injected volume: 10 µL			

Equipment control and data acquisition were performed with UNIFI v1.8.2 software (Waters, USA). Finally, external calibrations of mass and drift time curves were conducted weekly with the "Major Mix IMS/Tof calibration kit" directly purchased to Waters, prepared and infused at a flow rate of 20 µL/min for both positive and negative calibrations as well as CCS calibration. For internal lock mass stability, a Leucine-Enkephalin solution (50 ng/mL) in ACN:H_2O (50:50 v/v) at 0.1 % HCOOH was pumped at 10 µL/min through the lock-spray needle and measured every 30 seconds, with a scan time of 0.4 seconds. Leucine-enkephalin, in positive ($[M+H]^+$, *m/z* 556.2771) and negative mode ($[M-H]^-$, *m/z* 554.2615) was used for recalibrating the mass axis during the injection and to ensure a robust accurate mass along the time. Samples were injected in both positive and negative ionization modes.

Data processing

Data from the VION instrument (*.uep, UNIFI) were imported to *Progenesis QI* (NonLinear Dynamics, Newcastle, UK) and peak picked automatically during import process. The first ten samples (QC samples) were injected to stabilize the column using the last one as reference for retention time alignment in Progenesis QI. Samples were divided into groups (*QC*, *Herb* and *Tobacco*) in the "*Experiment Design Setup*" and final data was exported to Excel format containing for each feature, its *m/z* ratio, RT, CCS and abundance. Additional data, such as charge state, isotope distribution fit, peakwidth or p-value from ANOVA analysis can also be exported.

Features abundance was log2 normalized to avoid heteroscedasticity and Pareto scaled before importing for statistical analysis.

Statistical analysis

Multivariate analysis was carried out with EZ-Info (Umetrics, Sweden). Data were first analysed by Principal Component Analysis (PCA) in order to ensure that normalization have been correctly performed for each table (RP and HILIC in both positive and negative ionization mode). When QC samples were observed together in the center of the plot, all four data tables were joined into a single file for further feature selection.

Finally, OPLS-DA was performed in order to extract a smaller group of markers to differentiate between herbs and tobacco.

Elucidation workflow

Accurate masses for the most significant ions from OPLS-DA were retrieved from the feature table. Tandem mass spectrometry experiments (MS/MS) were also performed in order to obtain a richer product ion spectrum of the selected markers for their elucidation. Using the accurate mass, elemental composition calculator (MassLynx, Waters) and with the help of mass spectra databases (Metlin, Massbank), in-silico fragmentation web resources (MetFrag) and searching in general chemical database (Chemspider), compounds were tentatively elucidated.

When reference standards were available, they were purchased and injected to confirm their identity in the samples. When unavailable, RT and CCS values were predicted using the RT (Bade et al., 2015) and the CCS prediction tools (Bijlsma et al., 2017) in order to provide extra confidence to the tentatively elucidated compound.

Results and discussion

Experimental setup

Synthetic cannabinoids are generally mixed with or sprayed onto selected group of herbal components such as *Canavalia Maritima, Nymphaea Alba, Scutellaria Nana, Leonotis Leonurus, Zornia Latifolia, Nelumbo Nucifera, Leonurus Sibiricus, Althaea Officinalis, Rosa Canina, Turnera Diffusa, Verbascum Thapsus, Astragalus, Calendula Officinalis* or *Trifolium Pratense* (Ogata, Uchiyama, Kikura-Hanajiri, & Goda, 2013) (European Monitoring Centre for Drugs and Drug Addiction (EMCDDA), 2015). These herbs were purchased and 0.5 g of each one were rolled in a cigarette (by triplicate). Then, cigarettes were introduced into an ENVI-Carb® SPE cartridge previously conditioned as indicated in the section. Cartridges were introduced in a vacuum system and cigarettes were lighted, being "smoked" continuously by the system as mentioned in the "Sample treatment" section from Materials and Methods.

All cigarettes were rolled with the same cigarette paper and no filter was employed, in order to avoid the introduction of new variables to the experiment. Sample treatment becomes a key part in the metabolomics process, as the more sample treatment applied the more compounds can be lost. For this reason, the strategy was to trap the smoke produced by these herbs in specific SPE cartridges based on activated carbon, which are considered optimal for trapping volatile compounds (Fredes 2016)

UHPLC-IMS-QTof MS based metabolomics. Data processing and multivariate analysis.

As a test of the benefits and requirements that four dimension data (RT, CCS, *m/z* ratio and integrated area) has in metabolomics field, smoke extracts were injected in a VION instrument. Both positive and negative ionization modes (to cover a wide range of acidity/basicity) in two different chromatographic mechanisms, RP (C_{18}) and HILIC (in order to cover a wide range of polarities) were acquired.

Data was exported from UNIFI in **.uep* extension. Progenesis QI, provided by Non-linear dynamics is, at this moment, the only data processing software able to interpret this file format. This program, which is very user-friendly, guides the user to import data, selecting a reference sample (in this case a QC sample) in order to correct retention time. This QC is equivalent to the use of external standards in target analysis, with the main benefit that represents all the samples in the set.

Progenesis QI also allows revising the deconvolution process and identifying compounds by searching them in online databases as Chemspider. Finally, the user is able to transcript data to EZ-info, implemented in the software, or to export data for statistical analysis using other statistical programs. After export process, feature labels were manually modified to "Mxxx.xTyyy.yCzzz.z_AAA", being xxx.x the nominal Mass, yyy.y the retention Time (in seconds), zzz.z the CCS value and AAA the chromatographic column and ionization mode (RP+, RP-, HI+ or HI-). Data abundances were log2 transformed and Pareto scaling was applied.

PCA was performed for each dataset, observing that QC samples are grouped in the centre of the plot (**Figure 1** , RP pos as an example), which ensures that differences between groups are not going to be caused by differences in instrumental processing but for real changes in compound occurrences. After that, all four data tables were joined into a single file and statistically analysed by Orthogonal Partial Least Squares – Discriminant Analysis (OPLS-DA).

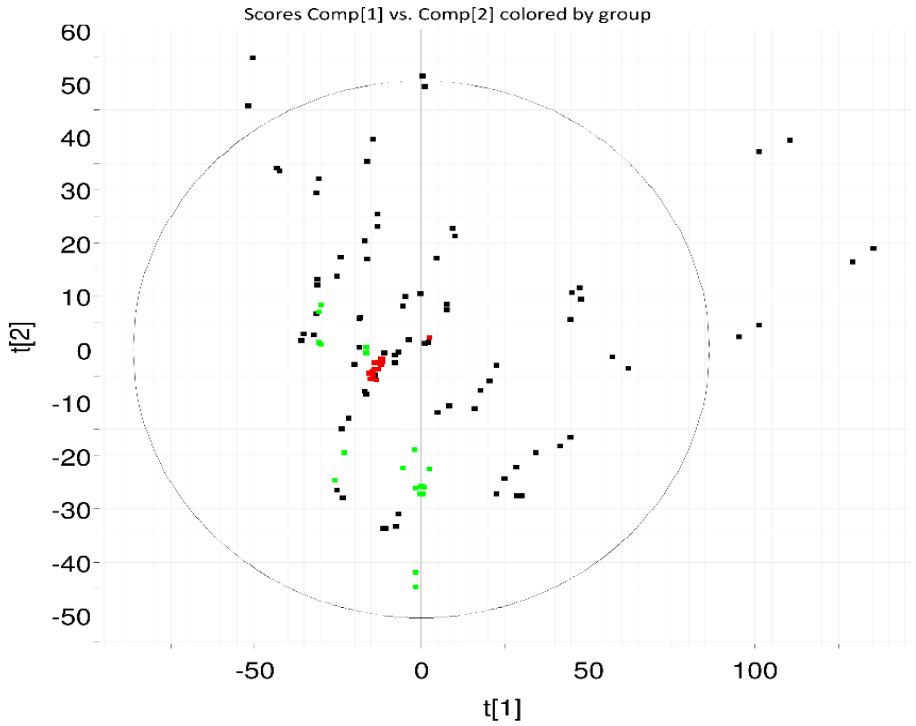


Figure 1: PCA of RP in positive ionization mode. Red points (QC samples) are joined in the center of the plot. Green points are tobacco samples and black points herb samples.

OPLS-DA was carried out to highlight the most discriminative markers between tobacco and herb samples. From the total of 18671 features, only few showed high significant difference between the groups, as can be observed in the generated S-Plot (**Figure 2**). This plot assign a number between -1 and 1 to each feature (ion) called P[corr]. The most discriminative ions between both groups are in the extreme parts of the plot (up-right or down-left part), which has P[corr] nearer to -1 or 1, depending if they are higher in tobacco (1) or herb (-1).

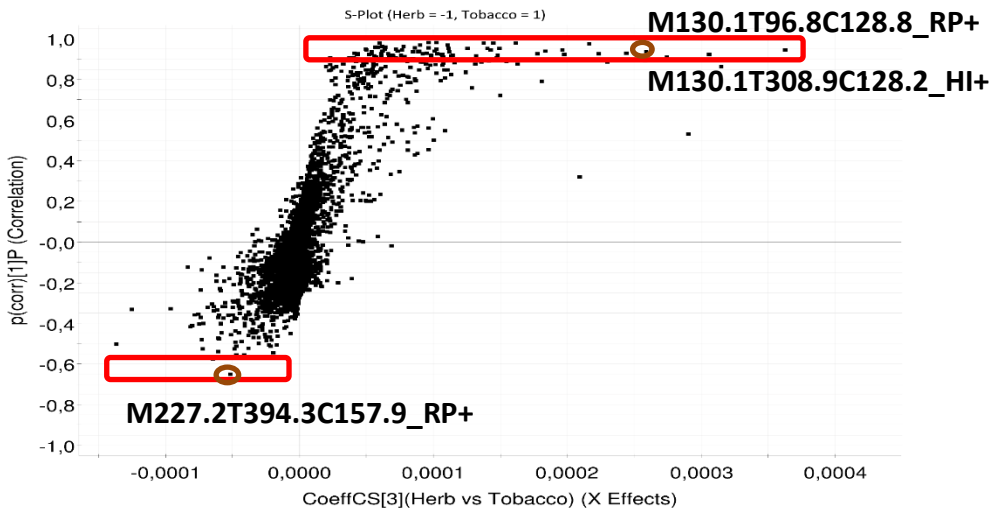


Figure 2: S-Plot representation of the OPLS-DA. Red circles show the most discriminant ions in both groups. Up are compounds higher in tobacco and down compounds high in herb samples. Ions rounded in Green corresponds to selected markers.

In these end parts of the plot a reduced group of ions were observed, with more compounds higher in tobacco than in herbs. This is probably because of the homogeneous composition within tobacco samples compared to the 14 different herbs, with heterogeneous compositions.

Figure 2 shows an ion with high P[corr] (labelled M227.2T394.3C157.9_RP+, P[corr]=0.653, marker 2), while two features were observed higher in tobacco with the same exact mass and closely similar CCS value (M130.1T96.8C128.8_RP+ and M130.1T308.9C128.2_HI+, with P[corr]=0.953 and 0.942, marker 1), corresponding to the same compound detected in both HILIC and RP analysis.

Although Marker 2 showed a smaller P[corr] than Marker 1, probably because of the heterogeneity of the herbs group, the fact that a single compound was highlighted in such a variable group constituted by so different samples, makes its use very promising.

The ratio between selected markers could be more significant to differentiate groups, than absolute markers responses. As can be observed in **Figure 3B**, all tobacco samples had a ratio

(Marker 2/Marker 1) lower than 2 (mean=0.39 ± 0.36), while all the herbs showed a considerably higher ratio (mean=79 ± 53) none below 2, showing the discrimination power of this combination. For this reason, both markers were finally selected for their elucidation.

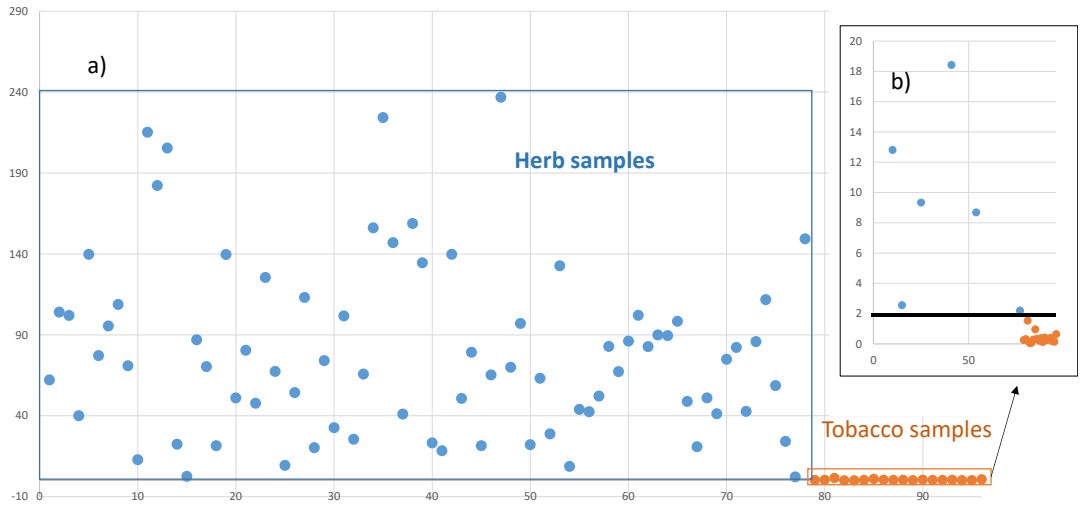


Figure 3: a) Comparison of the ratio Marker 2/Marker 1 in trapped smoke between herb and tobacco samples. b) zoom in the ratio between both compounds with Y-axis between 0 and 20.

Elucidation process and RT/CCS prediction for selected markers

In order to elucidate selected markers, MS^E spectra were analysed in order to obtain a plausible candidate taking profit of the mass accuracy of both parents and fragments. Comparing **Figure 4** (drift time filtered against no filtered), the benefits of being all the ions ion-mobility separated before the MSE acquisition can clearly observed. The drift time alignment between both low and high energy spectra cleans both spectra from interfering ions appearing at the same retention time but not at the same drift time, observing only the fragments, almost product ions, derived from the specific "precursor" ion., rendering a quasi MS/MS spectrum. However, the collision energy ramp, used as a compromise, was not enough to generate abundant product ions in the high energy spectrum (**Figure 4**) where no product ions were observed. For this reason, conventional product ion scans were performed at different fixed collision energies (CE), (10, 20, 30 and 40 eV) in order to obtain richer spectra for elucidation processes.

The most likely elemental composition for marker 1 with accurate mass *m/z*=130.06500, retention time of 1.48 min and CCS value of 128.8 $Å^2$, was found to be $C_9H_8N^+$ (error: -0.7 mDa) (Figure 5a). The predominant product ions (PI) at 20 eV CE (**Figure 5b**) were PI1, with *m/z* 103.05237 ($C_8H_7^+$, error: -2.4 mDa), assigned to the neutral loss of HCN (27.01263) and PI2, with *m/z* 77.03846 ($C_6H_5^+$, error: -0.6 mDa), assigned to a consecutive neutral loss of C_2H_2 (26.01391). PI2 showed a double bond equivalent of 4.5, pointing out the presence of four unsaturations, which only can correspond to a benzene ring for this empirical formula. In this sense, after searching in Metlin, *quinoline* or *isoquinoline* were the most likely tentative identifications as well as *ethenyl-benzonitrile*. However, the last one was discarded in front of the others because it should loss C_2H_2 chain directly from the parent ion, showing a product ion at *m/z* 104.0494 which was not observed in the spectrum

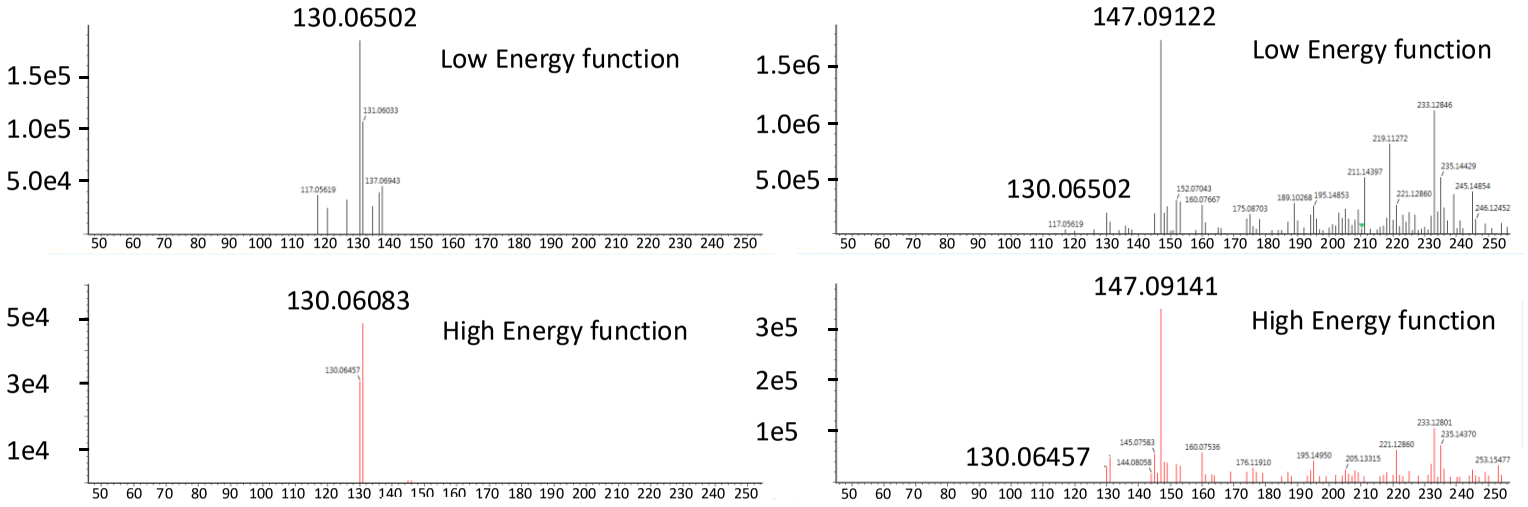


Figure 4: a) MS^E for Marker2 with (left) and without (right) filtering by drift time.

In order to gain more confidence in the tentative identification, RT and CCS predictors were employed. RT predictor estimated an RT of 4.1 min while CCS tool predicted a value of 127.9 $Å^2$. As can be observed, CCS value fits well with the experimental one (deviation -0.7%), while RT showed considerable deviation of +177%. However this deviation could be explained by the fact that the mobile phase was acidified around pH 3, and quinoline(Pk_a=4.5) was protonated in solution showing a lower RT. Although RT prediction can be affected by different parameters, ion mobility separation conditions are more controlled. Therefore, based on the higher confidence obtained due to CCS prediction, the quinoline reference standard was purchased. MS/MS experiments were performed at 10, 20, 30 and 40 eV collision energies. As can be seen in **Figure 5**, product ions of Marker 1 (**Figure 5a, 5b** and **5c**) perfectly matched with tandem mass spectra of quinoline standard (**Figure 5d, 5e** and **5f**). Furthermore, RT 1.51 min and CCS 128.8 $Å^2$ showed small deviations -2 % and 0 %, respectively. All this information lead to the unambiguously identification of marker 1 as quinoline.

Marker 2 showed an accurate mass *m/z*=227.15362, retention time of 6.69 min and CCS value of 157.9 $Å^2$. Three possible elemental compositions fit within a 3 mDa mass error ($C_{15}H_{19}N_2^+$, error: -1.2 mDa, $C_{12}H_{21}NO_3^+$, error: 1.5 mDa and $C_{10}H_{19}N_4O_2^+$, error: 2.8 mDa). **Figure 6** shows two product ions with *m/z* 183.09091, *m/z* 184.09998 at 20 eV CE, in addition product ions with m/z 168.06691, 156.08049 , 130.0668 , 103.05578 and 77.03957 were shown at 30 eV CE. The same product ions than quinolone, as well as its protonated molecule m/z 130.0668, were observed. This indicated that Marker 2 could have similar backbone structure as quinoline. Furthermore, the presence of PI3 (assigned to a neutral loss of C_3H_8, error: 0.1mDa) and PI4 (assigned to a radical loss of $C_3H_7^{\cdot}$, error: -1.2 mDa) pointed out to the loss of an isopropyl moiety bonded to a nitrogen, which was lost as HCN to PI7 (error: 0.5 mDa). No matches were found in Metlin, therefore a MetFrag search was performed selecting Chemspider as reference database. A total of 14 compounds were obtained as possible candidates with score upper to 0.9, but the most scoring compounds (Chemspider ID: 47250370, 50494419 and 52240184) contained the quinoline backbone structure. In order to achieve more confidence in the tentative elucidation, RT and CCS was predicted with the machine-based ANN predictors published in the literature by our group (Bade et al., 2015; Bijlsma et al., 2017). Experimental CCS value was 157.9 $Å^2$ while predicted CCS value was 155.8 $Å^2$ (1.3% error), 155.1 $Å^2$ (1.8 % error) and 156.1 $Å^2$ (1.1% error) for Chemspider ID: 47250370, 50494419 and 52240184,

respectively, while RT predicted was 6.03 min for three candidates, which is near to the experimental value of 6.67 min (9% error). As we can observe, predicted CCS values gave much lower errors than predicted RT, which are more affected to experimental changes (column aging, specific material, etc.) than ion mobility. In this case, the higher error observed for RT than CCS could be explained for a slightly different reversed phase material used during ANN model building. However, as in both cases a travelling wave ion mobility separator was used, the deviations in CCS predictions are expected to be smaller and a higher confidence might be expected from ion mobility data than retention time data. Hence, Marker 2 was tentatively identified as the isomer candidate with the lowest CCS deviation (Chemspider ID 52240184, see also **Figure 6**).

In view of the findings and to sum up, Marker 1 was confirmed with reference standard as quinoline and Marker 2 was tentatively elucidated as *N-Isopropyl-3-(8-quinolinyl)-2-propen-1-amine*.

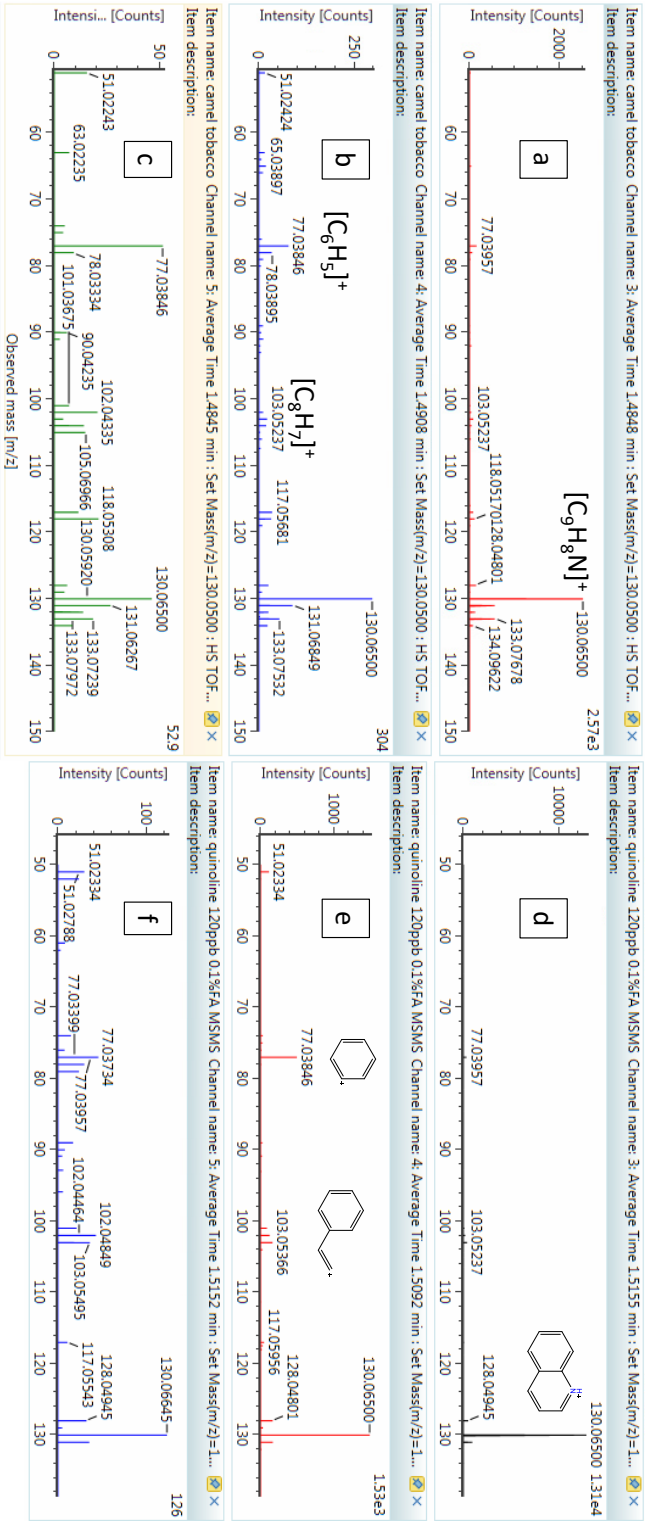


Figure 5: MSMS experiments for Marker 1 at 10 (a), 20 (b) and 30 (c) eV collision energies and for purchased quinoline standard at 10(d), 20 (e) and 30 (f) eV collision enegies

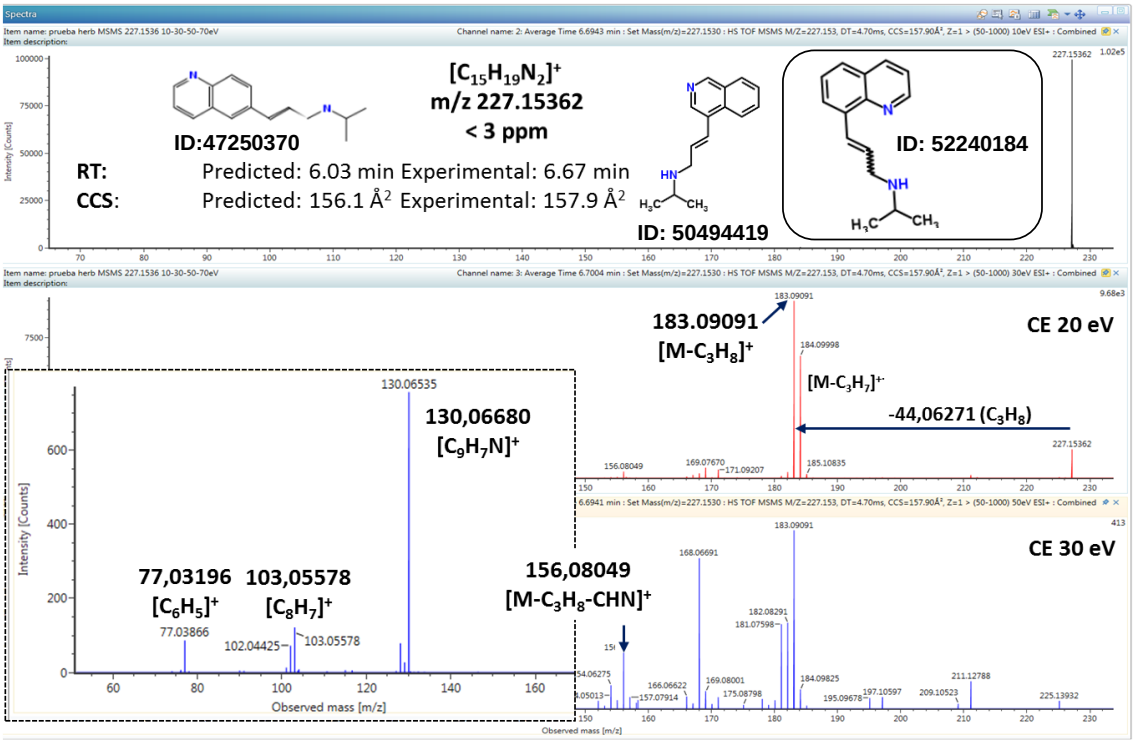


Figure 6: Experimental spectra obtained for Marker 2 with MS/MS experiments at 10, 20 and 30 eV CE. A proposed structure is showed as well as their predicted and experimental RT and CCS values.

Conclusions

The study confirms the capabilities of the state-of-the-art UHPLC-IMS-QTof-based metabolomics to obtain valuable biomarkers. The usefulness of performing metabolomics with UHPLC-IMS-QTof instruments have been pointed out, as tentatively elucidated markers, not available to purchase in the market, have been confidently confirmed with CCS predicted values, which gives a confirmation for our tentative elucidations (tandem mass spectra and predicted collisional cross section).

The robustness of CCS value prediction for "confirmation" purposes have been shown. An innovative way to highlight and elucidate unknown compounds in "poorly known matrices" have also been evaluated. Progenesis QI have also shown its potential to perform data mining with extra dimensions in data (collisional cross section information), being an friendly-user tool to perform data treatment.

Samples generated a really heterogeneous group, but the extended sensitivity of VION instrument coupled to the high selectivity of UHPLC made possible to found and highlight a couple of biomarkers, which ratio have been demonstrated as a good tool to differentiate tobacco and herbal samples used to spice.

The employed ratio between both markers, which varies from a mean of 79 for herbs to 0.39 to tobacco (more than 100 times higher), makes them a perfect signalling of herb composition, probably even in herb-tobacco mixes. The developed method can ensure the presence of spiced products when any targeted compound have been found. Now, its applicability to biofluids or wastewater analysis is under investigation.

Acknowledgments

Lubertus Bijlsma acknowledges NPS-Euronet (HOME/2014/JDRUG/AG/DRUG/7086), co-funded by the European Union, for his post-doctoral fellowship. Ruben Gil-Solsona acknowledge the financial support from SCORE-COST action ES1307 "Sewage biomarker analysis for community health assessment" for their Short Term Scientific Mission (STSM) grant (COST-STSM-ES1307-150916-080342). The authors acknowledge the financial support from Generalitat Valenciana (Group of Excellence Prometeo II 2014/023) and from the Ministerio Español de Economía y Competitividad (Project CTQ2015-65603-P).

References

Astarita, G., (2009). Applications of ion mobility MS in metabolomics. *Advanced LC-MS Applications in Metabolomics*, 94–109.

Bade, R., Bijlsma, L., Miller, T. H., Barron, L. P., Sancho, J. V., & Hernández, F. (2015). Suspect screening of large numbers of emerging contaminants in environmental waters using artificial neural networks for chromatographic retention time prediction and high resolution mass spectrometry data analysis. *Science of the Total Environment*, *538*, 934–941.

Beale, D. J., Karpe, A. V., McLeod, J. D., Gondalia, S. V., Muster, T. H., Othman, M. Z., … Joshi, D. (2016). An "omics" approach towards the characterisation of laboratory scale anaerobic digesters treating municipal sewage sludge. *Water Research*, *88*, 346–357.

Bijlsma, L., Bade, R., Celma, A., Mullin, L., Cleland, G., Stead, S., … Sancho, J. V. (2017). Prediction of Collision Cross-Section Values for Small Molecules: Application to Pesticide Residue Analysis. *Analytical Chemistry*, *89*(12), 6583–6589.

Bijlsma, L., Gil-Solsona, R., Hernández, F., Sancho, J.V. (2018). What about the herb? A new metabolomics approach for synthetic cannabinoid drug testing. *Analytical and Bioanalytical Chemistry410, 5107-5112*.

Cevallos-cevallos, J. M., Etxeberria, E., Danyluk, M. D., & Rodrick, G. E. (2009). Metabolomic analysis in food science : a review. *Trends in Food Science & Technology*, *20*(11–12), 557–566.

Couto, M., Barbosa, C., Silva, D., Rudnitskaya, A., Delgado, L., Moreira, A., & Rocha, S. M. (2017). Oxidative stress in asthmatic and non-asthmatic adolescent swimmers—A breathomics approach. *Pediatric Allergy and Immunology*, *28*(5), 452–457.

D'Atri, V., Causon, T., Hernandez-Alba, O., Mutabazi, A., Veuthey, J.-L., Cianferani, S., & Guillarme, D. (2017). Adding a new separation dimension to MS and LC-MS: What is the utility of ion mobility spectrometry? *Journal of Separation Science*.

Dettmer, K., Aronov, P. A., & Hammock, B. D. (2007). Mass spectrometry-based metabolomics. *Mass Spectrometry Reviews*, *26*(1), 51–78.

European Monitoring Centre for Drugs and Drug Addiction (EMCDDA). (2015). *EMCDDA | European Drug Report 2015*.

Hagan, N., Goldberg, I., Graichen, A., St. Jean, A., Wu, C., Lawrence, D., & Demirev, P. (2017). Ion Mobility Spectrometry - High Resolution LTQ-Orbitrap Mass Spectrometry for Analysis of Homemade Explosives. *Journal of the American Society for Mass Spectrometry*, *28*(8), 1531–1539.

Kaufmann, A., Walker, S. (2018). Comparison of linear interscan and interscan dynamic ranges of Orbitrap and ion-mobility time-of-flight mass spectrometers. *Rapid Communications in Mass Spectrometry, 31(22), 1915-1926*.

Ogata, J., Uchiyama, N., Kikura-Hanajiri, R., & Goda, Y. (2013). DNA sequence analyses of blended herbal products including synthetic cannabinoids as designer drugs. *Forensic Science International*, *227*(1–3), 33–41.

Raro, M., Ibáñez, M., Gil, R., Fabregat, A., Tudela, E., Deventer, K., … Pozo, Ó. J. (2015). Untargeted Metabolomics in Doping Control: Detection of New Markers of Testosterone Misuse by Ultrahigh Performance Liquid Chromatography Coupled to High-Resolution Mass

Spectrometry. *Analytical Chemistry*, *87*(16), 8373–8380.

Schymanski, E. L., Jeon, J., Gulde, R., Fenner, K., Ruff, M., Singer, H. P., & Hollender, J. (2014, February 18). Identifying small molecules via high resolution mass spectrometry: Communicating confidence. *Environmental Science and Technology*. American Chemical Society.

Zhang, X., Kew, K., Reisdorph, R., Sartain, M., Powell, R., Armstrong, M., … Reisdorph, N. (2017). Performance of a High-Pressure Liquid Chromatography-Ion Mobility-Mass Spectrometry System for Metabolic Profiling. *Analytical Chemistry*, *89*(12), 6384–6391.

Zhou, Z., Tu, J., Xiong, X., Shen, X., & Zhu, Z. J. (2017). LipidCCS: Prediction of Collision Cross-Section Values for Lipids with High Precision to Support Ion Mobility-Mass Spectrometry-Based Lipidomics. *Analytical Chemistry*, *89*(17), 9559–9566. 5

CAPÍTULO V

DISCUSIÓN Y CONCLUSIONES

V.5: Discusión

Como ya se ha comentado en la introducción general (Capítulo I), cualquier fase es crucial a la hora de obtener resultados correctos que se ajusten a lo deseado. El tratamiento de las muestras es por tanto clave a la hora de analizar el mayor número posible de compuestos con una mínima perdida de información. Además de ello, la fase de tratamiento de datos debe ser aplicada y comprobada para así poder tener un buen conjunto de datos analíticos que resulte fiel a las concentraciones reales de las moléculas. Los modelos estadísticos utilizados deben ser también correctamente seleccionados, aplicados y validados para así obtener un modelo discriminante fiable y robusto.

En este capítulo, se van a discutir en mayor profundidad los pasos seguidos en nuestra aproximación metabolómica que resultan claves a la hora de obtener buenos resultados.

V.1. Diseño experimental

V.1.1. Tratamiento de muestra

Debido a la estrategia utilizada en el desarrollo de los trabajos de esta Tesis Doctoral (metabolómica no dirigida o *untargeted*), donde no se realiza ninguna selección previa de compuestos, el diseño experimental debe comportar un análisis completo de las matrices con el objetivo de analizar la mayor parte de compuestos posibles.

Como ya se comentó en la introducción general, las matrices de alimentos pueden resultar complejas de analizar en un modo no dirigido, ya que nunca sabemos la cantidad de información que perdemos al seleccionar los métodos de extracción debido al elevado número de compuestos con diferentes características (polaridad, etc). Sin embargo en muestras de suero el rango de polaridades se reduce bastante, aunque pueden aparecer también compuestos de características diversas debido a la base acuosa de la muestra. Por otra parte, en los artículos del capítulo VI, los compuestos podrán incluso no aparecer en bases de datos, ya que son producidos por la combustión de hierbas. De este modo, cada uno de los artículos ha sido desarrollado con un tratamiento de muestra diferente.

En los artículos científicos 1 y 2 el tratamiento de muestra que se siguió, al tratarse de una matriz muy estudiada como es el suero, los tratamientos utilizados se basaron en la bibliografía existente (*Malkar, A. et. al., 2013*). En base e ello, el tratamiento de las muestras consistió en una deproteinización, para intentar evitar pérdidas de compuestos con acetonitrilo. Posteriormente, dependiendo del tipo de cromatografía (fase reversa o HILIC), se realizó un cambio de solvente, intentado preservar como ya se ha dicho el máximo número de compuestos.

En el artículo científico 3, en el que se trabaja con aceite de oliva, la matriz líquida se dividió en dos partes a procesar. La primer alícuota se trató empleando disolventes polares (metanol) por lo que contendría los compuestos más polares del aceite. La otra alícuota, simplemente se diluyó con disolventes apolares (*n*-butanol) en donde se determinarón los compuestos de menor polaridad.

Este tratamiento se planteó debido a la composición de los aceites de oliva, donde los ácidos grasos se encuentran en tanto por ciento (más de 1000 mg/Kg), mientras que los compuestos polares (fenoles, lignanos, flavonoides,…) se miden en unos pocos mg/Kg (*Fuentes et al., 2017*). La dilución 1:10 en butanol resultó suficiente para medir, siempre de un modo no dirigido, los ácidos grasos presentes en los aceites en forma de triglicéridos, diglicéridos,…, donde se obtuvo una cantidad considerable de *features*. La extracción con metanol, por su parte, facilitó la medida de compuestos con menor concentración. Estas dos aproximaciones resultaron en una serie de marcadores de clasificación con los que la diferenciación de los aceites de oliva resultó correcta, como se puede observar en la sección III.2.

Sin embargo, cabe recordar que un método de autentificación de alimentos metabolómico se desarrolla con el principal objetivo de crear un método dirigido (*targeted*), generalmente con equipos de baja resolución (QqQ). Dicho esto, puede resultar tedioso a la par que caro desarrollar un método en el que cada muestra deba de ser extraida hasta en 4 ocasiones para poder establecer el origen de la misma. Por este motivo, en el artículo científico 4, a pesar de que también se había llevado a cabo la extracción de los compuestos lipídicos presentes en las almendras, finalmente no se continuó y, el modelo propuesto se desarrollo en base a los compuestos obtenidos por medio de una extracción polar.

La extracción utilizada se basó en un procedimiento publicado por nuestro grupo de trabajo, donde el método de extracción es muy robusto para una gran variedad de matrices alimentarias (*Beltrán et al., 2013*). A diferencia del aceite de oliva, la matriz de almendra es sólida, por lo que es necesaria una etapa previa de homogeneización de la muestra para reducir el tamaño de partícula al mínimo posible y así asegurar una extracción reproducible.

El disolvente utilizado fue una mezcla ACN:H_2O (80:20) con 0.1% de ácido fórmico (HCOOH). De este modo, se extraen compuestos muy polares a la vez que otros con un rango de polaridad intermedia (gracias al ACN). Además, la presencia de HCOOH puede facilitar la protonación de compuestos básicos de la matriz, aumentando su polaridad, para poder ser determinados en esta extracción.

Se optó por utilizar este método de extracción, considerado como general, evaluando que el número de *features* detectadas no fuera muy pequeño. Cabe destacar que con una estrategia de no dirigida es difícil llevar a cabo una optimización del disolvente de extracción, ya que un número elevado de *features* detectados no tiene una relación directa con una mejor separación entre los grupos. Sin embargo, un número elevado de *features* sí puede ir ligado a analizar el/los compuesto/s que realmente puedan ser útiles para la clasificación.

La cantidad de compuestos necesarios para diferenciar las almendras extranjeras de las almendras españolas, así como de las variedades nacionales, se obtuvieron de una mezcla entre los compuestos ionizados en modo positivo y negativo. Estos compuestos, que en instrumentos de triple cuadrupolo pueden ser analizados en el mismo experimento en modo *polarity-switching* (*Yuan et. al., 2012*), no requieren diferentes tipos de cromatografía, por lo que un único análisis es necesario para cada muestra, reduciendo enormemente el tiempo y costo del mismo. El análisis *lipidómico* de las almendras se encuentra en proceso en estos momentos, por lo que no se pueden dar resultados sobre la validez o no de esta fracción para la discriminación.

En el artículo científico 5, donde se trabajó con cafés procedentes de diferentes regiones de Colombia, se dio un paso más. Además de obtener diferentes fracciones, se ampliaron los tipos de cromatografía que se utilizaron. La extracción con ACN:H_2O (80:20), 0.1% HCOOH fue escogida debido a la similitud entre las almendras y el café verde en cuanto a compuestos polares, pero

además se realizó una extracción de la parte lipídica con el fin de comparar todos los rangos de polaridades y escoger de entre ellos el que mejor discrimina las diferentes matrices.

Para la extracción de la parte lipídica se utilizó un disolvente más apolar (acetona), debido a la facilidad que tiene ésta para evaporarse, lo que facilita la fase de cambio de disolvente posterior. En el caso del aceite de oliva, al tratarse de una matriz con más del 80% de compuestos lipídicos (triglicéridos,…), el disolvente escogido fue n-butanol, a pesar de que, por motivos cromatográficos, los picos de los compuestos menos retenidos se podrían ensanchar a la entrada de la columna, y presentar tiempos de retención menos reproducibles. Sin embargo, en una matriz como el café verde donde entre el 15 y el 17% de los compuestos presentes son considerados lípidos (https://www.coffeechemistry.com/chemistry/lipids/lipids-in-coffee), resulta interesante realizar el cambio de disolvente para adecuarlo al inicio del gradiente cromatográfico. Por este motivo la acetona se evaporó a vacío, redisolviendo el residuo seco con ACN, utilizado también como fase móvil al inicio del gradiente para este tipo de compuestos. Al ser compuestos bastante apolares, estos quedaban retenidos en la columna a pesar del inicio con 100% orgánico, viéndose picos sin distorsión.

En cuanto al capítulo IV, como ya se ha mencionado en la introducción, la estrategia establecida ha sido propuesta para ser aplicada en cuatro matrices no invasivas diferentes, que son el humo producido por la combustión, el exhalado pulmonar contaminado por éste, la saliva y la orina del consumidor. El exhalado pulmonar, debido a la reducida concentración de los compuestos en el mismo y a que se encuentra en fase gas, debe ser analizado por medio de instrumentos de análisis directo (*Martinez-Lozano Sinues, et. al., 2013*). Sin embargo, el humo producido por la combustión de hierbas base es más fácil de atrapar y la concentración de los compuestos generados en el mismo es de esperar que sea mucho mayor, pudiendo ser atrapados en cartuchos de extracción en fase sólida, como se ha propuesto en el artículo científico 7.

Para ello, el humo se hizo pasar a través de los cartuchos seleccionados, que fueron finalmente eluídos y, tras un cambio de disolvente, inyectados en el equipo UPLC-ESI-IMS-QTOF MS. Como se ha comentado en la introducción, probablemente un tratamiento de muestra tan restrictivo puede no ser la primera opción en la mayoría de experimentos desarrollados con fines

metabolómicos. Sin embargo, en una muestra en fase gas resulta ser una buena opción. Se escogió un cartucho que pudiera atrapar un amplio rango de compuestos, como es el carbón activado, donde se recogen tanto compuestos polares como apolares, ya que su capacidad de extracción se basa en la adsorción de los mismos sobre la superficie de las partículas de carbón, pudiéndose desorber posteriormente con disolventes adecuados.

Por otro lado, la saliva puede funcionar como un absorbente natural para compuestos que se encuentran en el humo, por lo que su uso en este tipo de estrategias puede ser beneficioso. Los compuestos atrapados en la misma, que se compone principalmente de agua (99% aprox., *Álvarez-Sánchez B. et al., 2012*), se encontrarán probablemente dentro de un rango de elevada polaridad, por lo que esta matriz resultaría perfecta para la cromatografía de líquidos. Sin embargo, como se expone en la literatura (*Malkar, A. et. al., 2013*), debido a la baja concentración esperada de los compuestos, es aconsejable un paso adicional de preconcentración además de deproteinizar la muestra.

Finalmente, la orina se postula como una matriz capaz de revelar el consumo de estas sustancias a medio-largo plazo, ya que el resto de matrices (humo y exhalado, pero principalmente la saliva) probablemente no presenten ventanas de detección demasiado amplias. En este sentido, resultará muy ventajoso realizar el mismo experimento en las cuatro matrices seleccionadas disponiendo de una herramienta útil a lo largo del tiempo a la par que robusta. La visión conjunta de los resultados en todas las matrices nos facilitará asegurar que los compuestos provienen del humo de las hierbas.

V.1.2. Análisis cromatográfico

En cuanto al análisis cromatográfico, para los artículos científicos del capítulo II se ha escogido la utilización de dos tipos diferentes de cromatografía de líquidos, la cromatografía de fase reversa y la cromatografía HILIC. El objetivo de utilizar dos tipos diferentes de cromatografía es que, puesto que en estos artículos no se pretendía obtener un modelo de clasificación sino estudiar los compuestos que se veían modificados, podemos observar los comportamientos de un rango más amplio de polaridades. Además, debido a la naturaleza no dirigida de los experimentos, no es posible crear un gradiente específico, por lo que se realiza un gradiente lineal, esperando que la

mayoría de los compuestos de la muestra presenten un comportamiento adecuado y puedan ser detectados correctamente. Del mismo modo en los artículos científicos del capítulo IV no existe un conocimiento a priori del tipo de compuestos que se puedan encontrar en él, por lo que se decide utilizar el mismo tipo de cromatografías y gradientes para los experimentos.

Sin embargo, para los artículos científicos en los que se trabaja con alimentos, al tener una mejor previsión del tipo de compuestos presentes en los extractos, se utilizaron cromatografías que pudieran separar satisfactoriamente las familias de compuestos presentes en los mismos.

En el artículo científico 3 se constata la dificultad que podemos encontrar al trabajar con matrices altamente grasas como el aceite de oliva. En este artículo se discute la necesidad de realizar un lavado entre inyecciones debido al contenido lipídico tan elevado de las muestras.

Se realizó un análisis con patrones de referencia de ácidos grasos libres, mono-, di- y triglicéridos para intentar desarrollar una correcta separación tanto entre grupos como entre los compuestos del mismo grupo. Se comprobó cómo añadiendo una pequeña cantidad de agua en la fase móvil A (mayoritariamente metanol) se obtenía una mejor resolución entre picos cromatográficos de ácidos grasos y monoglicéridos. Esta mejor separación entre los mismos resulta necesaria para distinguir entre las respuestas de, por ejemplo, ácido oleico y *monooleoyl glicerol* . Una pequeña fragmentación en la fuente de ionización del monoglicérido generaría iones de ácido oleico que no se podrían distinguir del ácido oleico libre que hubiese en la muestra, en la caso de coelución cromatográfica.

Sin embargo, esta pequeña adición de agua implicó que una pequeña parte de los triglicéridos quedasen adsorbidos en la válvula de inyección, ya que el inicio del gradiente contenía muy poca proporción de n-butanol en la fase móvil (B). Este efecto memoria, podría afectar la composición observada de la muestra siguiente. Por ello, entre cada una de las inyecciones se realizó un gradiente sin inyección comenzando por 50% de n-butanol en la fase móvil B, que permitía prácticamente eliminar de la válvula estos interferentes (3-4% de efecto memoria residual) y poder llevar a cabo una medida correcta de las muestras, no afectando a la diferencia entre grupos, ya que la proporción retenida se mantenía constante entre inyecciones.

En cuanto al artículo científico 5, se utilizó una columna HILIC para obtener una mejor separación cromatográfica de los compuestos polares en la matriz. El procedimiento de extracción fue el mismo para el análisis por RP y por HILIC, aunque el rango de compuestos analizados por ambas técnicas se amplió, obteniendo una imagen más universal de la composición de los granos de café verdes. Finalmente, como ya se ha comentado anteriormente, solo uno de los métodos cromatográficos se utilizó en el desarrollo del modelo de clasificación, a pesar de que el tratamiento de datos se realizó conjunto. En este caso, se seleccionaron solamente compuestos de la parte RP en modo de ionización negativo debido a que estos presentaron mucha mejor capacidad de predicción a la hora de discriminar los grupos.

V.2. Tratamiento de datos

El tratamiento de datos seguido en todos los artículos científicos excepto el artículo 7 fue el mismo, con software libre como es el XCMS. El método de *peak picking* utilizado fue *CentWave*, con el que se obtuvo un buen número de *features* en todos los casos. Los tiempos de retención fueron alineados correctamente, sin observar ninguna variación superior a 10 segundos previa al alineamiento, lo que demuestra la estabilidad de las columnas UHPLC. En el paso que se observó una mayor desviación fue a la hora de realizar la normalización de las señales, debido posiblemente a la caída de señal que provoca una inyección muy prolongada de muestras que, sin embargo, fue corregida tras la misma.

En los artículos científicos 1 y 2, el tratamiento de datos a pesar de resultar necesario, no provocó grandes modificaciones en la matriz, ni en la alineación de tiempos de retención ni en la normalización de las señales. Esto fue así debido al relativamente pequeño número de muestras inyectadas (alrededor de 35 en ambos casos). Esta cantidad de muestras no pareció ser suficiente para provocar una caída de señal importante en el instrumento, por lo que los datos obtenidos experimentalmente no requirieron de un esfuerzo especial en este caso.

En el artículo científico 3, como se puede observar en el material suplementario (**Sección III.6, Artículo científico 3, Figure S4**), las muestras QC presentan una caída de señal

considerable a lo largo de la secuencia, que se corrige con la normalización aplicada. Las muestras QC se inyectaron inicialmente 10, donde no se observa una caída de señal, posteriormente se inyectaba una muestra QC cada 10 muestras (3 horas y 20 minutos). Al haber un total de 90 muestras, la secuencia de inyección duró aproximadamente 33 horas, por lo que la caída observada se puede considerar normal, debido a la posibilidad de que el cono de extracción acumule matriz no volatilizada haciendo que la sensibilidad del equipo decaiga. Sin embargo, como se puede observar, la normalización soluciona este problema, obteniendo un conjunto de datos mejor ajustado a la realidad química, por medio de la normalización a la mediana. Este método ha demostrado ser muy útil en metabolómica (*K. A. Veselkov et al., 2011*), ya que cuando se tienen matrices con una composición muy parecida el compuesto que representa la mediana tendrá un nivel igual en todas las muestras.

En el artículo científico VII, sin embargo, se ha trabajado con *Progenesis QI* como software de procesamiento de datos. Los datos se adquirieron con información en cuatro dimensiones (*m/z* ratio, tiempo de retención, movilidad iónica e intensidad) que XCMS, que había sido el software de elección hasta el momento, no era capaz de procesar. Adicionalmente, no existe hasta la fecha ningún programa informático que permita transformar datos en extensión *.uep* a *.cdf*, para poder emplear programas gratuitos. Por este motivo, el único programa que es capaz de extraer la información de los datos crudos, *Progenesis QI*, fue el empleado en esta ocasión.

Este software resulta muy intuitivo y permite realizar de un modo completamente guiado, los mismos procesos que con el XCMS, obteniendo finalmente una lista de datos integrados con sus cuatro dimensiones. Entre ellos, como se puede comprobar en la **Figura V.1**, se encuentra la importación de los datos, alineación de los tiempos de retención, *peak picking*, identificación de los posibles compuestos por medio de *ChemSpider* o estadística multivariante con *EZ-Info*.

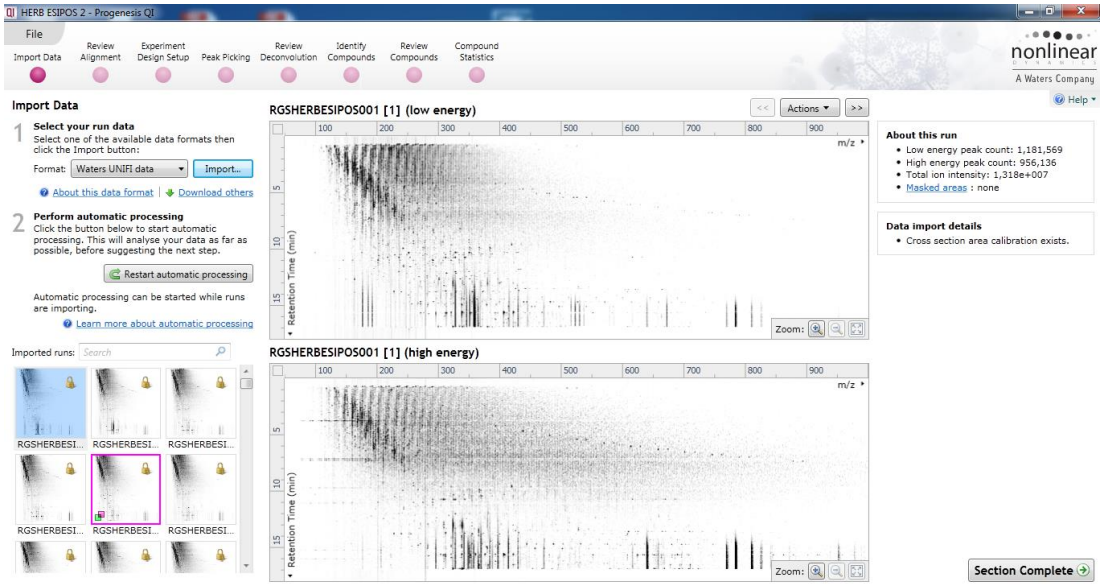


Figura V.1: Captura de pantalla de Progenesis QI.

Sin embargo, a pesar de poder realizar la estadística con el mismo Progenesis QI, se decidió exportar los datos y valorar si se debía realizar un paso extra de normalización antes de extraer la información relevante, como se comenta a continuación.

V.2.1. Uso de la normalización

Como ya se ha comentado en la introducción, la normalización es muy útil cuando se tienen unas condiciones cambiantes a lo largo de la inyección de todo el conjunto de muestras, ya que la acumulación de suciedad de la interfase o posibles variaciones ambientales, como temperatura pueden condicionar la respuesta del instrumento y generar diferencias donde *a priori* no debería haberlas. Como ya se ha expuesto en otros capítulos de esta Tesis Doctoral, la normalización es muy útil en sets de muestras donde a pesar de que las matrices son prácticamente idénticas (como en el caso de los sueros de peces o de muestras del mismo alimento), se observa mediante las muestras QC, que son utilizadas como patrón externo, que existe una caída importante en la señal analítica a pesar de que se inyecte la misma muestra (ver **Figura I.12**, Capítulo 1).

Sin embargo, en el artículo científico VI la caída observada no es tal, como se puede comprobar en la **Figura V.2**, donde los QC apenas sufren variación a lo largo de la secuencia de inyección,por lo que se decidió no aplicar ninguna otra normalización de las muestras y seguir trabajando con estos datos.

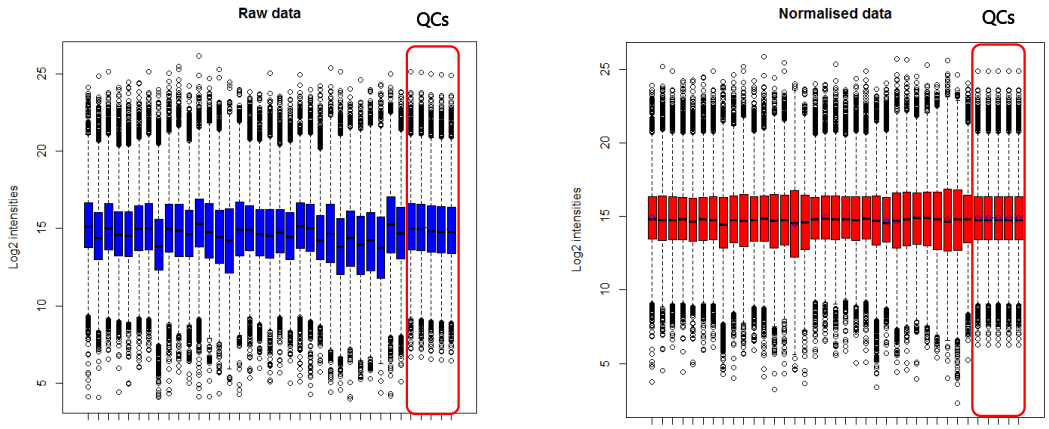


Figura V.2: Boxplots de todos los features hallados para las muestras de saliva. A la izquierda, datos sin normalizar, a la derecha, datos normalizados. Enmarcardas en rojo, muestras QC.

En el artículo científico 6 las muestras, a pesar de pertenecer a sujetos que habían fumado hierbas muy heterogéneas, estaban en una matriz de saliva, donde los compuestos endógenos generaban muestras muy parecidas, como se puede apreciar en los *Boxplot* de la **Figura V.2**. Sin embargo, en el artículo científico 7, Las muestras de humo de hierbas y tabaco difieren enormemente entre ellas, encontrando compuestos en común pero también otros hallados únicamente en un grupo o incluso en una hierba en concreto, por lo que la heterogeneidad de las muestras es mucho mayor, como se observa en la **Figura V.3**.

Como se puede apreciar, las muestras de tabaco (a la derecha de los QC) presentan una cantidad de compuestos (número de features) muchísimo mayor y con mayores intensidades. En este caso, aplicar una normalización sería un error, ya que estaríamos equiparando señales que realmente son muy diferentes intentando corregir una "posible" deriva instrumental. Además, esta deriva no existe, ya que si se observa con detención las muestras QC, la mediana de sus señales así

como su dispersión son casi inapreciables antes de normalizar. Por ello se decidió en este caso proseguir sin aplicar la normalización a los datos.

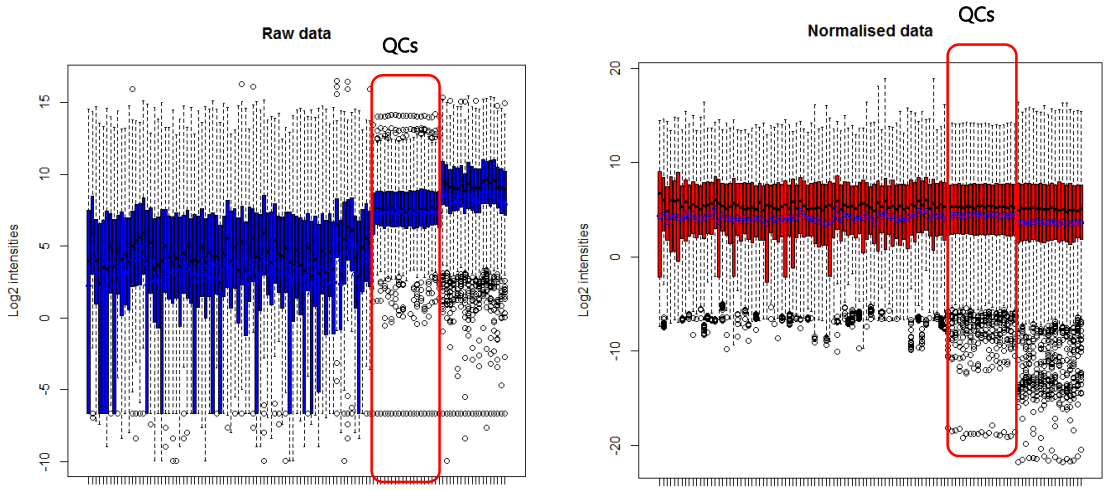


Figura V.3: Boxplots de todos los features hallados para las muestras humo en tabaco y hierbas. Izquierda, datos sin normalizar, Derecha, datos normalizados. Enmarcadas en rojo, muestras QC. Dentro de cada figura, las muestras de hierbas se encuentran a la izquierda de los QCs y las de tabaco a la derecha.

Sin embargo, en todos los experimentos se aplicó el logaritmo en base 2 para corregir la heteroscedasticidad y el *Pareto scaling*, ya que ninguno de estos afectan a la integridad de las muestras y, por lo tanto, no estaremos sobre-corrigiendo nuestros datos.

Finalmente, se puede concluir que el instrumento con el que se adquirieron las muestras, en este caso el VION IMS QTOF (Waters) presenta una deriva instrumental muy pequeña (**Figura V.3**) si se compara con el XEVO G2 QTof, también de Waters (**Figura V.2**), utilizado en los demás experimentos, ya que las muestras QC se observan perfectamente alineadas sin aplicar ningún tipo de normalización.

V.3. Análisis estadístico

En los artículos que componen esta Tesis Doctoral se intentó evaluar diferentes modelos tanto de selección de variables como de autentificación, por lo que esta apartado se divide a su vez en tres subsecciones: *Modelos de selección de variables*, *Modelos de discriminación y autentificación* y *Modelos univariantes*

V.3.1. Modelos de selección de variables

En el artículo científico 3 se utilizó el OPLS-DA como modelo de selección de variables. Este modelo estadístico discriminante permite encontrar los marcadores individuales que, enfrentando dos grupos diferentes de muestras, están más correlacionados con un grupo que con otro. De este modo, se enfrentó cada una de las zonas geográficas individualmente al resto, obteniendo un marcador específico para cada una de ellas. Estos marcadores, que posteriormente fueron elucidados, se validaron con el método SVSM (descrito en la introducción general, capítulo I). De manera que se aseguró que todos los marcadores seleccionados resultaban necesarios en la discriminación.

Por otro lado, en los artículos científicos 4 y 5, se utilizó el valor de VIP proporcionado por PLS-DA para reducir las variables. Este método de reducción se utilizó, inicialmente, para obtener un número reducido de posibles marcadores (unos 50 compuestos) que generase un modelo válido (el cual se comprobó por medio de la validación). Posteriormente se iban eliminando aquellos *features* que presentasen un VIP inferior a 1 hasta obtener un modelo insatisfactorio (que presentara un porcentaje de acierto inferior al 90%) y se omitía este último, manteniendo el último que proporcionaba un acierto mayor al 90%.

De este modo, tanto en el modelo de clasificación de aceites de oliva (con 12 marcadores), en el modelo de las almendras (con 5 marcadores para el país, y 20 marcadores para la clasificación varietal) o en el modelo de los cafés (con 13 marcadores para la clasificación zonal) se llegó a un reducido grupo de marcadores que permitía establecer una correcta clasificación.

Finalmente, tras evaluar ambos métodos, se puede establecer el modelo de selección de VIP como más fácil de aplicar. Simplemente con unos pocos pasos en el tratamiento estadístico es

posible obtener un reducido grupo de marcadores. Además, al ser generados por PLS-DA, es casi instantáneo realizar la validación inicial de las muestras (por medio de una validación cruzada). Sin embargo, al intentar obtener los marcadores por medio del OPLS-DA, se debe generar un modelo para cada grupo enfrentado al resto y, posteriormente, filtrar estos marcadores manualmente y validar la utilidad de estos por medio de un PLS-DA, siendo un proceso mucho más laborioso y con una incertidumbre mayor a la hora de conocer si los marcadores empleados son satisfactorios. Aunque a la vista de los resultados ambos procesos proporcionan modelos de clasificación correctos, finalmente se decidió por utilizar el modelo de clasificación por VIP puesto que es mucho más simple y menos tedioso.

V.3.2. Modelos de discriminación y autentificación

Con el objetivo de crear el modelo de autentificación se han utilizado también diversos modelos estadísticos.

En el artículo científico 3 (**Capítulo III.1**), se utilizó el modelo de clasificación discriminante de SPSS, utilizando como variable de inclusión la lambda de Wilks, que es un modelo de una prueba de hipótesis multivariante. Este modelo permite generar funciones discriminantes que expliquen la máxima varianza posible y que se utilizarán para establecer a que grupo pertenece cada muestra nueva.

En el artículo científico 4 (**Capítulo III.2**), se utilizó PLS-DA como modelo discriminante, donde, al igual que el modelo de SPSS, las muestras generaban un modelo estadístico que explicara un % de la varianza y las nuevas muestras se introducían en el modelo generando un *output* con la clasificación.

Sin embargo, a la hora de generar el modelo discriminante para el artículo científico 5 (**Capítulo III.3**), se observó en la bibliografía que recientemente se había publicado un artículo cuestionando la validez de PLS-DA en autentificación de alimentos (*Rodionova et al., 2016*). Como se comenta en el citado artículo, los modelos de varios grupos obtienen muy buenos resultados clasificando nuevas muestras de grupos ya incluidos en el modelo, ya que son análisis dirigidos. Sin embargo, análisis basados en métodos multivariantes no dirigidos (como PCA) funcionan mejor a la

hora de no clasificar incorrectamente muestras pertenecientes a grupos no introducidos en la generación del modelo, muestras denominadas "*muestras alien*".

Por este motivo, se decidió comparar la utilidad de ambos modelos discriminantes frente a muestras *alien*, como se puede observar en el artículo científico 5. PLS-DA resultó inconsistente a la hora de clasificar muestras pertenecientes a grupos no introducidos en el modelo (muestras *alien*) al generar falsos positivos, mientras que DD-SIMCA por su parte no generó ningún falso positivo. Por este motivo se llega a la conclusión de que DD-SIMCA es un modelo estadístico adecuado para crear modelos discriminantes aplicables a muestras reales, mientras que PLS-DA ha demostrado su versatilidad y potencial a la hora de observar las diferencias entre grupos y mostrarlas, aunque no es un buen método a la hora de generar modelos estadísticos discriminantes.

De este modo, los sets de datos de los artículos científicos 3 y 4 se re-utilizaron para generar los modelos discriminantes por DD-SIMCA. Se puede observar cómo, a modo de ejemplo, los modelos generados por DD-SIMCA en aceites de oliva (**Figura V.4**) y almendras (**Figura V.5**) funcionan correctamente frente a muestras *alien*, al igual que ha sucedido en el artículo científico V, mientras que el porcentaje de acierto se mantuvo igual.

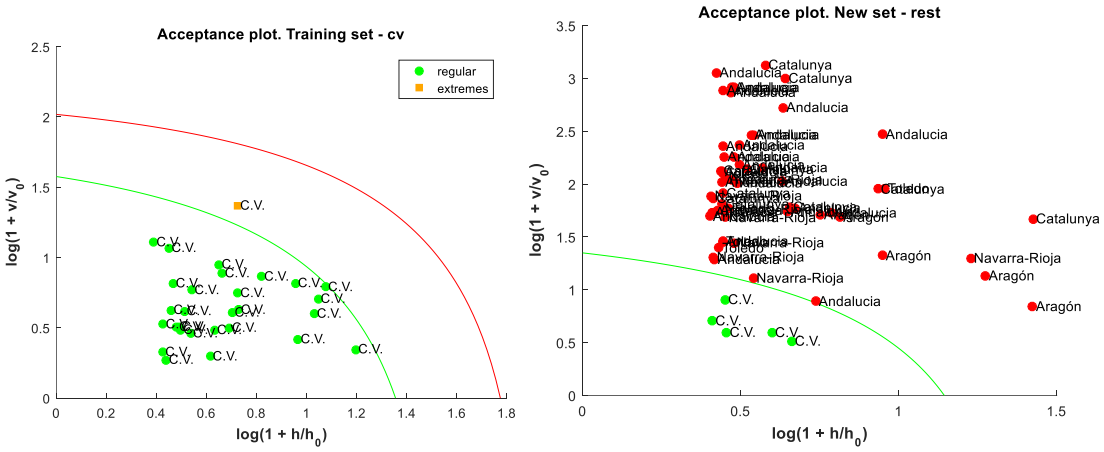


Figura V.4: Modelo DD-SIMCA generado para aceites de oliva de la comunidad valenciana. A la izquierda se muestra el modelo, a la derecha la validación.

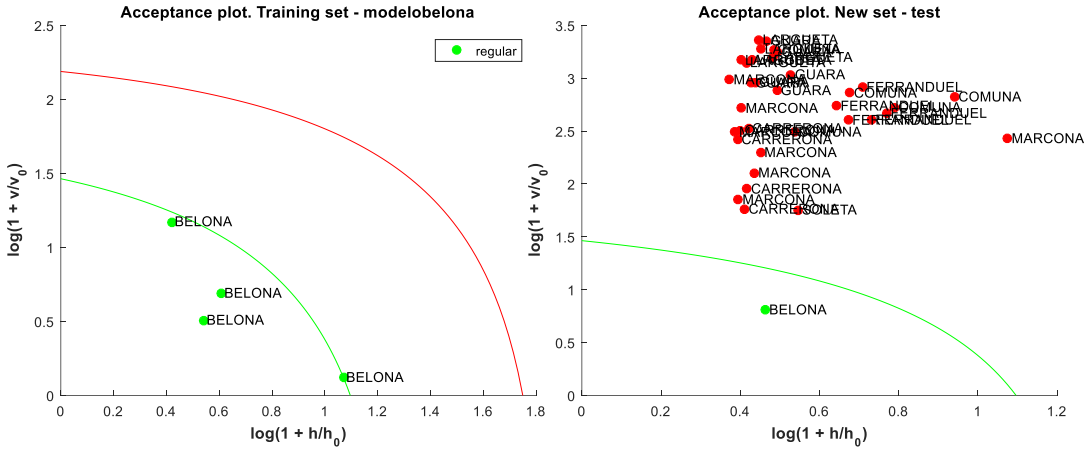


Figura V.5: Modelo DD-SIMCA generado para almendras de la variedad *Belona*. A la izquierda se muestra el modelo, a la derecha la validación.

V.3.3. Modelos Univariantes

En los artículos científicos del capítulo II se utilizaron, además de modelos estadísticos multivariantes (utilizados para observar de manera general el comportamiento de las muestras y sus posibles agrupaciones), modelos de estadística univariante con el fin de comprobar como los compuestos que afectan a las agrupaciones observadas de forma multivariante también son significativamente diferentes de modo univariante.

Para ello se utilizó el análisis de la varianza ANOVA, con el objetivo de observar las diferencias entre los grupos. Estas diferencias se pueden observar en las tablas de los artículos científicos 1 y 2, donde cada uno de los compuestos era estudiado individualmente para generar patrones comunes y lograr una mejor comprensión de las mencionadas variaciones.

A modo de ejemplo, podríamos utilizar las carnitinas ligadas a los ácidos grasos, compuestos íntimamente relacionados con la movilización de ácidos grasos en el organismo. Como podemos ver en el artículo científico 1, un ayuno drástico en los peces aumenta la cantidad de estos compuestos, ya que la grasa disponible se moviliza de forma abundante y esto se refleja en los análisis y en las tablas biométricas de este mismo artículo. Sin embargo, en el artículo científico 2, podemos comprobar como la biometría de los peces no es significativamente distinta

independientemente de la dieta con la que se les alimente. Por tanto, como los animales tienen disponibilidad de alimento, este tipo de compuestos no se ven alterados y no aparecen en las tablas (**Tabla II, artículo científico 1**). Sin embargo, en el caso de las fosfocolinas, se observa una disminución general en los animales en ayuno en el artículo 1 mientras que en el artículo 2 (**Capítulo II.2**) solo se observa la sustitución de ácidos grasos de origen marino por otros de origen vegetal. Estas conclusiones se alcanzan gracias al análisis específico e individual que se consigue a partir del estudio univariante.

V.4. Validación de los resultados

Una vez los marcadores fueron seleccionados para generar el modelo, se procedió a su validación. Esta parte resulta clave a la hora de proporcionar marcadores que sean capaces de discriminar correctamente, no solo en el conjunto de muestras, sino en toda la población a lo largo del tiempo. Por este motivo, los modelos de autentificación de alimentos se validaron siempre siguiendo tres fases tal y como se aconseja en la literatura (*Riedl et al., 2015*): i) *cross-validation* del modelo, ii) validación con muestras del mismo año y iii) validación con muestras de otro año (*system challenge*).

Sin embargo, aparece un problema colateral al intentar comparar muestras inyectadas con una variación temporal tan alta. Esta contrariedad está relacionada con la fluctuación de la sensibilidad instrumental en función del estado del detector del espectrómetro de masas, que puede afectar significativamente a la señal obtenida para los compuestos marcadores.

Esta traba, que aparentemente se podría solucionar con la adición de un patrón interno, no resulta tan fácil de afrontar a la hora de generar y validar el modelo metabolómico. Esto es debido a que, por una parte, los compuestos que van a ser utilizados como marcadores son completamente desconocidos *a priori*, por lo que no se puede pre-seleccionar un patrón interno válido para los futuros marcadores.

Por otro lado, el posible uso de un patrón externo está sujeto a los mismos problemas que el patrón interno. El uso de muestras QC como una especie de patrón externo también se descarta,

al ser éstas una mezcla de las muestras que generan el modelo. Por ello, pueden sufrir modificaciones a lo largo del tiempo (muestras de dos años diferentes), a pesar de encontrarse congeladas, y tener una deriva de señal demasiado importante. Por este motivo, se decidió aplicar otro tipo de normalizaciones entre años para solucionar este problema.

En la bibliografía se encuentran diferentes modos de normalización para solucionar estos problemas en el cambio de la señal instrumental (*Di Guida et al., 2016*), como puede ser la normalización por suma, donde se divide cada valor por la suma del total (representado un % del total) o la "*Probabilistic Quotient Normalization*" (PQN), donde se genera un vector con cada una de las respuestas de las muestras. Posteriormente, al valor mínimo se le asigna 0 y al máximo 1, con lo que la respuesta de una nueva muestra es transformada en la escala dada por el vector. Todos los artículos del capítulo III se normalizaron aplicando una normalización PQN.

En el artículo científico 3, el valor medio del vector que definía cada marcador en las muestras que generaron el modelo se igualó al valor medio del vector de validación, asegurando que ambos vectores tenían el mismo punto medio (corrigiendo así la variación de señal instrumental), a pesar de que no se autoescaló este vector para situarlo entre 0 y 1 (**Figura V.6**). De este modo la validación aportó buenos resultados.

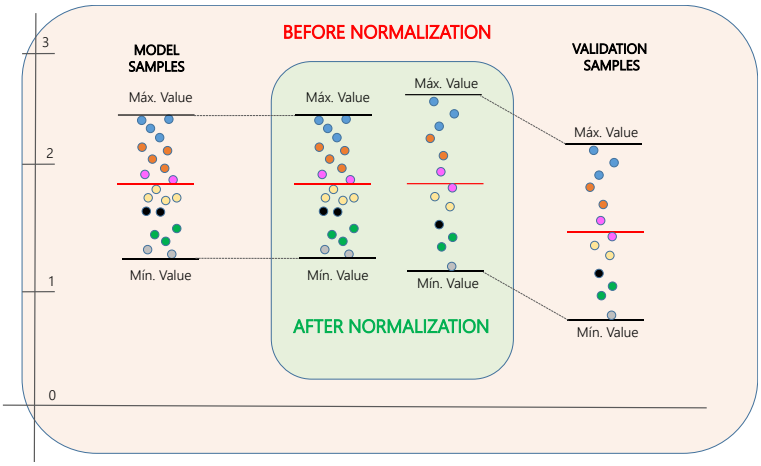


Figura V.6: Esquema de la normalización aplicada en el artículo científico 3

En el artículo científico 4, debido a que los compuestos utilizados como marcadores en los modelos tenían una señal muy elevada, no fue necesario aplicar ninguna normalización, ya que

las medias de ambos vectores (vector del modelo y vector de validación) se encontraban muy cercanas.

En el artículo científico 5, por su parte, se aplicó el mismo modelo de validación que en el artículo 3, pero en lugar de igualar las medias al valor del vector del modelo, el vector se autoescaló para igualar la desviación estándar debido a que se observó una mayor variabilidad entre las muestras (**Figura V.7**).

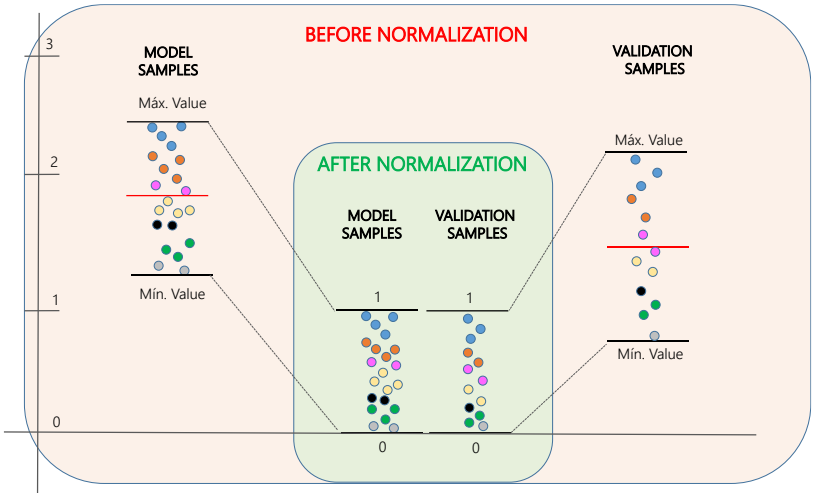


Figura V.7: Esquema de la normalización aplicada en el artículo científico 5

Una vez se obtuvo una correcta normalización entre secuencias de análisis, se procedió a la validación en lostres pasos propuestos anteriormente. De este modo se asegura que los marcadores observados como discriminantes son robustos a lo largo del tiempo.

La etapa final de todos los trabajos realizados en el capítulo III consistió en el desarrollo de un método dirigido (*targeted*), corregido con patrón interno, que debería ser validado con un número mucho mayor de muestras. Este método se podría convertir en el método de control de calidad de alimentos *premium* que la industria alimentaria podría incorporar con fines de control de la autenticidad de alimentos.

En cuanto al resto de artículos científicos, en los que el objetivo no era obtener un modelo de clasificación en sí, sino tratar de elucidar compuestos que nos permitan comprender e interpretar

mejor las agrupaciones preestablecidas (control-ayuno en el artículo 1, tipo de dieta en el artículo 2, tabaco-hierba en los artículos 6 y 7), no fue necesaria una normalización de este tipo ya que solamente se utilizó el lote de muestras a partir del cual se elucidaron los compuestos correspondientes.

V.5. Identificación y anotación

En este apartado se discutirá la necesidad de utilizar las técnicas más potentes a nuestro alcance para poder obtener una alta seguridad en la identificación de los marcadores seleccionados. Como ya se ha comentado anteriormente tanto en la introducción como en los artículos científicos, la obtención de modelos discriminantes gana mucho peso si es acompañada de la identidad molecular de los compuestos químicos involucrados.

V.5.1. Resonancia Magnetica Nuclear (RMN) y Espectrometría de Masas de baja y alta resolución

En los artículos científicos del capítulo II, principalmente en el artículo científico 1, se hace especial hincapié en la comparación entre los artículos publicados que trabajan con RMN y con espectrometría de masas. RMN presenta una universalidad y poder de elucidación al que la Espectrometría de Masas (MS) no puede llegar, dado a que analiza todos los átomos de un determinado elemento (hidrógeno, carbono,…) presentes en un compuesto aportando información estructural. Sin embargo, a pesar de que la espectrometría de masas requiere de compuestos ionizados, su sensibilidad y su capacidad de multidetección es muy superior a la de RMN.

En muestras de plasma, RMN es capaz de detectar alrededor de un centenar de compuestos a concentraciones elevadas, mientras que la MS ha demostrado en los artículos científicos 1 y 2 ser capaz de observar más de 10.000 iones diferentes en un rango de concentraciones amplio. Esta sensibilidad hace que MS proporcione una visión mucho más global de las muestras a estudiar y sea escogida para llevar a cabo los estudios de esta Tesis Doctoral.

Sin embargo la Espectrometría de Masas de baja resolución basa su poder de elucidación en la búsqueda de compuestos en librerías experimentales, ya que la resolución unidad no permite discriminar compuestos con diferencias de masa menores a 1 Dalton. Es por ello que para poder abordar la elucidación de compuestos desconocidos se requiere del poder de identificación que

proporciona la información de masa exacta generada por la espectrometría de masas de alta resolución (HRMS). Como ya es sabido, los diferentes elementos presentan variaciones pequeñas de masa exacta (NH_3 tiene una masa exacta de 18.0343 mientras que OH tiene una masa exacta de 18.0105), que son imperceptibles para un analizador de masas como el cuadrupolo pero que analizadores HRMS como el tiempo de vuelo separan sin ningún tipo de problema. Es por ello que el uso de éstos proporciona, no solamente una reducción de posibles estructuras químicas a la hora de elucidar los compuestos señalados por la estadística, sino una mejor asignación de las pérdidas que éste tiene en experimentos de masas en tándem.

En los artículos científicos del capítulo II, en los que el objetivo fue elucidar compuestos diferentes entre los grupos experimentales seleccionados, HRMS se muestra como una herramienta de extrema utilidad, ya que permite elucidar un gran número de compuestos, apoyado en el uso de bibliotecas de espectros de fragmentación como puede ser METLIN (https://metlin.scripps.edu/landing_page.php?pgcontent=mainPage) o human metabolome database (http://www.hmdb.ca/) y a la selectividad que ofrece la separación cromatográfica. Gracias a la combinación de ambas es posible obtener un buen conocimiento de los compuestos alterados, extendiendo el análisis a compuestos relacionados con los que ya han sido elucidados, en una especie de análisis *a posteriori*, como se ha hecho en el artículo científico 1 (*Refining process*).

Además de las bondades ya expuestas anteriormente, y debido al modo de análisis de los analizadores de tiempo de vuelo, todos los iones generados son analizados al trabajar en modo de espectro completo. Esta información permanece almacenada para cada inyección, por lo que es posible realizar en un futuro, un análisis retrospectivo sin necesidad de reinyectar las muestras. Esto resulta de utilidad si tras realizar el análisis se quieren observar una serie de compuestos diana que pueden resultar de interés, como se hizo en el artículo científico 2 con las vitaminas.

V.5.2. Movilidad iónica

En el artículo científico 7, se ha utilizado la movilidad iónica como información adicional a la hora de facilitar la elucidación de los marcadores. Con el uso de instrumentos UHPLC-ESI-IMS-QTOFMS, además del tiempo de retención de un compuesto en la columna cromatográfica o de su relación masa/carga, también se obtiene información de la sección de colisión transversal (Collision Cross Section, CCS). Este parámetro está relacionado con la forma del ión en el espacio mientras atraviesa un tubo de deriva con un flujo de gas inerte en contracorriente, donde un CCS mayor implica colisionar con más moléculas de gas y tardar más tiempo (drift time) en cruzar la celda. Esta dimensión de separación adicional se puede utilizar en varios sentidos.

El primero puede ser el de separar potenciales isómeros que tienen el mismo tiempo de retención y la misma relación *m/z*. Como se observa en la **Figura V.8**, un ión con relación *m/z* 295.1941 aparece, al mismo tiempo de retención, pero con dos valores de *drift time* diferentes (4.46 y 5.61 ms). La resolución por movilidad iónica de este instrumento es suficiente para que ambos picos se muestren bien resueltos, aunque el segundo tiene una intensidad muchísimo mayor que el primero.

Una segunda aplicación podría ser facilitar la elucidación de compuestos totalmente desconocidos, gracias al uso de herramientas de predicción, tanto del tiempo de retención (Bade et al., 2015) como, en particular, de la sección transversal de colisión, CCS (Bijlsma, Bade, et al., 2017). Se podría comparar el CCS medido para un ión, con los CCS predichos a partir de las estructuras químicas de los potenciales candidatos que todavía queden tras aplicar un filtrado por masa exacta, perfil isótopico y fragmentación plausible. Esta aproximación se ha aplicado satisfactoriamente en el artículo científico 7, donde la elucidación tentativa del marcador 2 se ha hecho tanto en base a su fragmentación en experimentos de MS/MS como a obtener unos valores predichos para el tiempo de retención y sección de colisión transversal dentro de los rangos de tolerancia validados (ver **Artículo científico 7, Capítulo IV.3**).

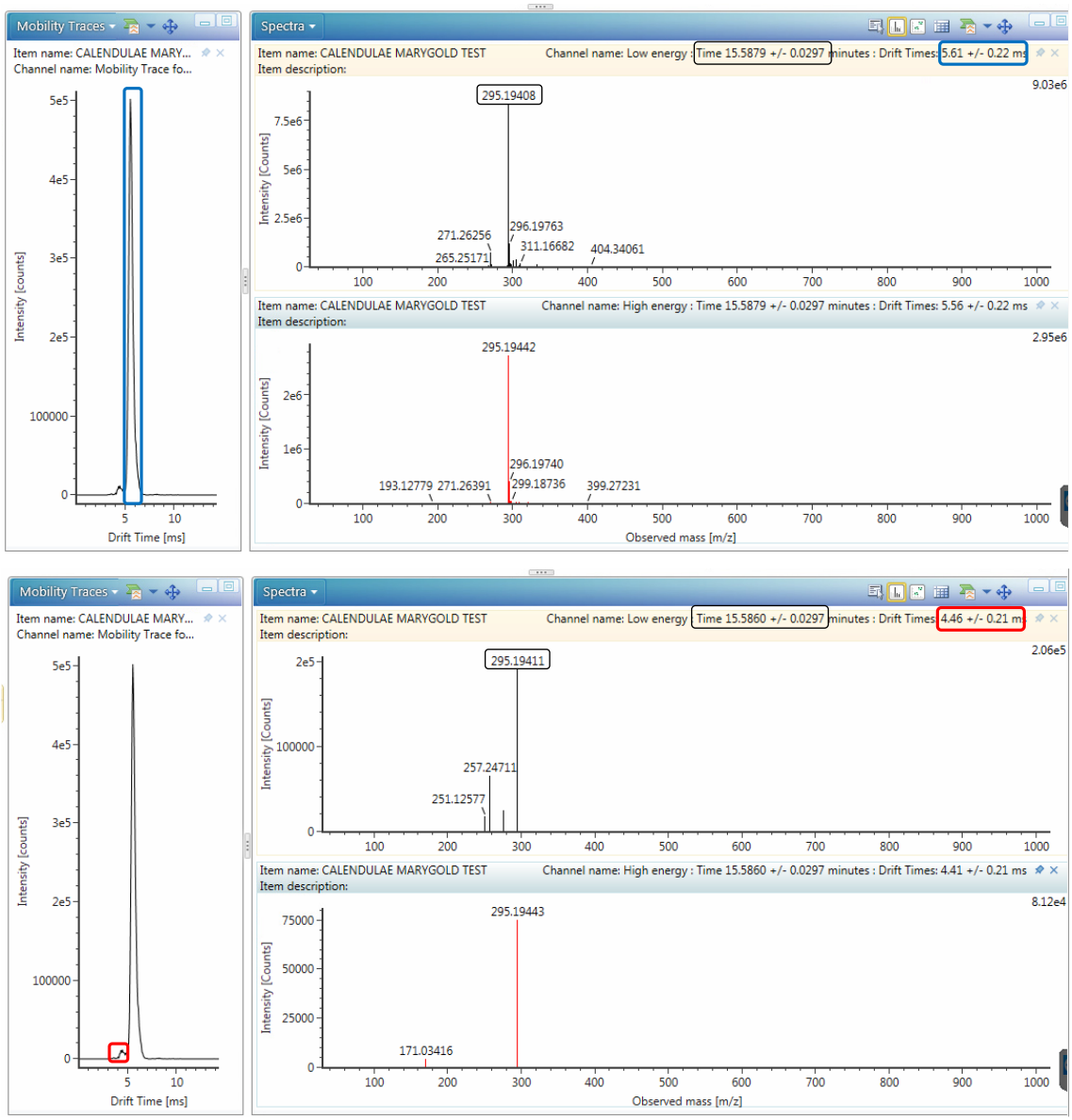


Figura V.8: Ión con m/z 295.1941 encontrado en una muestra de hierba correspondiente al artículo científico 7. Enmarcado en negro se observa la relación *m/z* y el tiempo de retención de ambos compuestos, con cuadro rojo y azul su correspondiente drift time así como su pico correspondiente en el mobilograma de la izquierda.

V.6. Interpretación biológica

Dependiendo de cuál sea el objetivo del experimento, la interpretación biológica juega un papel crucial a la hora de extraer conclusiones del trabajo. Es por ello que los artículos científicos del capítulo II, realizados en colaboración con expertos en nutrición y metabolismo animal, tienen una parte predominante basada en la interpretación biológica de los resultados obtenidos por la metabolómica.

Como se puede observar en estos artículos, los compuestos elucidados se agrupan dependiendo de su función biológica y esta interpretación acompaña a los resultados analíticos.

En el artículo científico 1, donde se realiza una prueba del poder que la HRMS tiene a la hora de extraer información de experimentos biológicos, se puede observar como hay una enorme cantidad de compuestos diferentes entre ambos grupos. De ellos se seleccionaron los que se habían visto modificados de una manera más significativa y a raíz de esto se interrelacionaron entre ellos, para poder dar una explicación general a las modificaciones individuales. De este modo se planteó que, como muchos de ellos estaban relacionados con ciclos metabólicos (como el ciclo de Meisters), pero no se habían elucidado todos los integrantes del ciclo, se estudiaría de una manera dirigida los compuestos relacionados para comprobar como el ciclo se había visto afectado, dando un mayor robustez a la interpretación funcional.

De los compuestos elucidados de una manera no dirigida, como el ácido piroglutámico, se obtuvieron las señales del resto (glutatión, glu-cys, glutamato) y se estableció como afectaba el ayuno a los compuestos relacionados con este ciclo.

De un modo semejante se trabajó en el artículo científico 2, donde los compuestos obtenidos de forma no dirigida se intentaron interrelacionar entre sí para darle un sentido biológico a los resultados. Sin embargo, la diferencia de planteamiento entre ambos experimentos es diametralmente opuesta. El primero busca poner a prueba la metabolómica con un experimento en el que las diferencias debían ser muy abultadas y como se ha observado en el artículo científico 1, la metabolómica basada en HRMS fue capaz de obtener una enorme cantidad de compuestos alterados y, a su vez, los que resultaban más robustos para definir un estado de malnutrición. Sin

embargo, en el artículo científico 2, el objetivo era demostrar que no había cambios entre las dietas utilizadas.

A pesar de que se observaron cambios, ninguno de ellos se encontraba relacionado con un estado de malnutrición, ya que el comportamiento de estos no era el mismo que en el artículo científico 1.

En lugar de ello, los cambios se debían a la composición de las dietas. En el artículo se vió como al sustituir una parte lipídica de origen marino (rica en ácidos omega-3,6…) por otra de origen vegetal (ácido palmítico, linoleico, oleico,…), los ácidos grasos circulantes en la sangre reflejaban estos cambios de composiciones, al igual que lo hacía la taurina (procedente de fuentes animales y no vegetales). En un análisis dirigido, para afianzar los resultados y conclusiones obtenidas se buscaron los niveles séricos de vitaminas, comprobándose como las únicas variaciones observadas eran en vitaminas sintetizadas por la microbiota intestinal que, dependiendo del tipo de dieta que el animal ingiere, se ve afectada.

Bibliografía

Álvarez-Sánchez, B., Priego-Capote, F., & Luque de Castro, M. D. (2012). Study of sample preparation for metabolomic profiling of human saliva by liquid chromatography–time of flight/mass spectrometry. *Journal of Chromatography A*, *1248*, 178–181.

Beltrán, E., Ibáñez, M., Portolés, T., Ripollés, C., Sancho, J. V., Yusà, V., … Hernández, F. (2013). Development of sensitive and rapid analytical methodology for food analysis of 18 mycotoxins included in a total diet study. *Analytica Chimica Acta*, *783*, 39–48.

Di Guida, R., Engel, J., Allwood, J. W., Weber, R. J. M., Jones, M. R., Sommer, U., … Dunn, W. B. (2016). Non-targeted UHPLC-MS metabolomic data processing methods: a comparative investigation of normalisation, missing value imputation, transformation and scaling. *Metabolomics*, *12*(5), 1–14.

Fuentes, E., Paucar, F., Tapia, F., Ortiz, J., Jimenez, P., & Romero, N. (2017). Effect of the composition of extra virgin olive oils on the differentiation and antioxidant capacities of twelve monovarietals. *Food Chemistry*, *243*, 285–294.

Malkar, A., Devenport, N. A., Martin, H. J., Patel, P., Turner, M. A., Watson, P., … Creaser, C. S. (2013). Metabolic profiling of human saliva before and after induced physiological stress by ultra-high performance liquid chromatography-ion mobility-mass spectrometry. *Metabolomics*, *9*(6), 1192–1201.

Riedl, J., Esslinger, S., & Fauhl-Hassek, C. (2015). Review of validation and reporting of non-targeted fingerprinting approaches for food authentication. *Analytica Chimica Acta*, *885*, 17–32.

Rodionova, O. Y., Titova, A. V., & Pomerantsev, A. L. (2016, April 1). Discriminant analysis is an inappropriate method of authentication. *TrAC - Trends in Analytical Chemistry*. Elsevier.

Veselkov, K. A., Vingara, L. K., Masson, P., Robinette, S. L., Want, E., Li, J. V., … Nicholson, J. K. (2011). Optimized preprocessing of ultra-performance liquid chromatography/mass spectrometry urinary metabolic profiles for improved information recovery. *Analytical Chemistry*, *83*(15), 5864–5872.

Wang, M., Rang, O., Liu, F., Xia, W., Li, Y., Zhang, Y., … Xu, S. (2018). A systematic review of metabolomics biomarkers for Bisphenol A exposure. *Metabolomics*, *14*(4), 45.

Yuan, M., Breitkopf, S. B., Yang, X., & Asara, J. M. (2012). A positive/negative ion-switching, targeted mass spectrometry-based metabolomics platform for bodily fluids, cells, and fresh and fixed tissue. *Nature Protocols*, *7*(5), 872–881.

V.2: Conclusiones

Los estudios llevados a cabo durante la realización de esta Tesis Doctoral han permitido evaluar el potencial de la instrumentación analítica más reciente para su uso en metabolómica, tanto para obtener datos representativos de las muestras analizadas como para su posterior elucidación. Las conclusiones principales que se pueden extraer de los trabajos llevados a cabo son:

- El uso de espectrómetros de masas de alta resolución es una gran elección a la hora de enfrentarse a la elucidación de compuestos desconocidos.
- La utilización de diferentes tipos de mecanismos en cromatografía de líquidos proporciona una visión mucho más amplia de las polaridades de los compuestos, lo que repercute positivamente en los resultados obtenidos en metabolómica no dirigida.

Además de esto, también aparecen conclusiones específicas de cada capítulo:

- El uso de estrategias dirigidas y no dirigidas en el estudio de muestras con el objetivo de obtener la mayor información química posible es recomendable. Si bien es cierto que las estrategias no dirigidas subrayan las diferencias que se observarían de un modo dirigido, además de muchas otras, la vasta cantidad de información con la que se trabaja hace muy complicada la extracción de toda la información deseable, por lo que un conocimiento de las muestras de antemano puede proporcionar información de un modo guiado que sería complejo de extraer de un modo no dirigido.
- La tecnología metabolómica basada en espectrómetros de masas de alta resolución con la capacidad de adquirir datos en modo *full-scan*, como el QTof, se ha mostrado como una buena herramienta a la hora de extraer información a posteriori de los datos ya inyectados para validar y asegurar la información biológica obtenida de los experimentos.
- La correcta elección de las técnicas estadísticas con el objetivo de crear modelos de autentificación resulta clave a la hora de enfrentarse a muestras con variaciones desconocidas. Por ello, modelos estadísticos como el PLS-DA resultan útiles a la hora de extraer información de los experimentos si bien a la hora de crear los modelos de

clasificación se debe acudir a otro tipo de modelos, como el DD-SIMCA, capaces de enfrentarse sin mucho error a muestras "alien", las cuales no se han tenido en cuenta en el diseño del experimento.

- A pesar de que, en el estudio no dirigido de las muestras a través de la metabolómica, la combinación de diferentes mecanismos cromatográficos y/o modos de ionización resultan muy interesantes, a la hora de crear un modelo dirigido conviene seleccionar compuestos de un mismo experimento para evitar así excesivas variaciones a la hora de crear los modelos estadísticos de predicción.
- La tecnología de movilidad iónica ha demostrado su potencial a la hora de enfrentarse a matrices desconocidas, como el humo producido en la combustión de hierbas, proporcionando una nueva dimensión de separación que, unido a herramientas de predicción de sección de colisión, proporcionan una mayor confianza en los resultados obtenidos.

V.3: Conclusions

The different experiments performed along this Doctoral Thesis have allowed to evaluate the newest analytical instrumentation potential for their use in metabolomics. Their use have permit to obtain both representative data from the analysed samples and richer spectra information for elucidation purposes. The main conclusions obtained from the works can be:

- The use of high resolution mass spectrometers is preferred to face the isolation and elucidation of unknown compounds.
- The employment of different kinds of liquid chromatographic mechanisms provides a wide polarity coverage of our unknown compounds. This fact also favourably impact in the results obtain from untargeted metabolomics.

Furthermore, other specific conclusions can be obtained from these works:

- The use of targeted as well as untargeted strategies in the same experiment, with the main objective of extracting the most interesting chemical information is highly recommendable. Despite untargeted strategies contains differences found in a targeted way, the vast amount of data to be handled in these untargeted way makes really difficult to extract all the desired information. In this sense, a previous knowledge of our samples in a guided way can help to obtain valuable information which can be difficult to obtain in a untargeted way.
- Metabolomics technology based in high resolution mass spectrometry with the ability of obtaining full-scan data, as can be QTof instrument, have emerged as a good tool to extract information in a retrospective way from previously injected data in order to validate and ensure biologic information obtained in our experiments.
- Choosing correctly the statistical model for authentication purposes becomes a clue in order to face samples with unknown variations. For this reason, statistical models as can be PLS-DA are really useful to extract as maximum information as possible from our data, but in order to create statistical models, the use of other statistical models as can be DD-SIMCA is preferred. This fact is because DD-SIMCA can face in a correct way the

inclusion in the model of "alien samples", which has not be taken into account in the experimental setup.

- Despite the use of different chromatographic mechanisms and/or ionization modes is highly recommended to extract information in an untargeted way, it is preferable to perform the targeted experiments only with information from one of them. It is because the selection of only one experimental data helps to avoid excessive variations to perform prediction models.
- Ion mobility technology have demonstrated its potential to face matrices with a reduced knowledge about their composition. This is the example of the smoke produced in the combustion of the herbs employed in spice products which provides a new separation dimension that coupled to predictors, provides a higher confidence in our results.

Sugerencias para futuros trabajos

Después de todo el trabajo desarrollado, se detallan a continuación posibles trabajos a desarrollar en la misma línea:

- Realizar estudios en la línea de los artículos científicos I y II, pero con otras matrices biológicas que permitan obtener un mayor conocimiento del estado nutricional del animal.
- Llevar a cabo la validación de los artículos científicos III, IV y V por medio de un espectrómetro de masas como un triple cuadrupolo, y validarlo de nuevo con muestras de otros años para asegurar la robustez de los resultados.
- Aplicar la estratégia seguida en los artículos científicos VI y VII a las otras matrices propuestas (orina, sangre…) con la finalidad de obtener los mejores marcadores del consumo de estas sustancias.
- Aplicar la estratégia metabolómica a otro tipo de drogas difíciles de detectar, como podría ser el caso de la *Burundanga* (Escopolamina).